

BEST AVAILABLE COPY

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
27 December 2002 (27.12.2002)

PCT

(10) International Publication Number
WO 02/103320 A2

(51) International Patent Classification⁷: G01N

(21) International Application Number: PCT/US02/18947

(22) International Filing Date: 14 June 2002 (14.06.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/298,918 18 June 2001 (18.06.2001) US
60/380,710 14 May 2002 (14.05.2002) US

(71) Applicant: ROSETTA INPHARMATICS, INC.
[US/US]; 12040 115th Avenue, N.E., Kirkland, WA 98034 (US).

(72) Inventors: DAi, Hongyue; 16814 118th Avenue, NE, Bothell, WA 98011 (US). HE, Yudong; 11410 NE 124th Street #148, Kirkland, WA 98034 (US). LINSLEY, Peter, S.; 2430 9th Avenue West, Seattle, WA 98119 (US). MAO, Mao; 12445 110th Lane, #Q302, Kirkland, WA 98034

(US). ROBERTS, Christopher, J.; 7019 Brooklyn Avenue NE, Seattle, WA 98115 (US). VAN'T VEER, Laura, Johanna; Brouwersgracht 192-G, NL-1013 HC Amsterdam (NL). VAN DE VIJVER, Marc, J.; Reguliersgracht 21, NL-1017 LJ Amsterdam (NL). BERNARDS, Rene; Koningsvaren 37, NL-1391 AD Abcoude (NL). HART, A., A., M.; Molenweide 44, NL-1902 CH Castricum (NL).

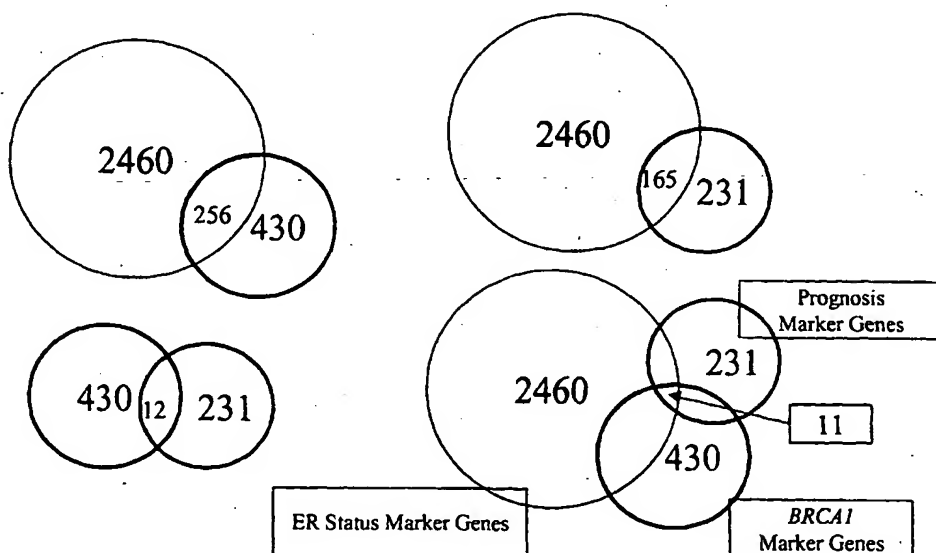
(74) Agents: ANTLER, Adriane, M. et al.; Pennie & Edmonds LLP, 1155 Avenue of the Americas, New York, NY 10036 (US).

(81) Designated States (*national*): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZM, ZW.

(84) Designated States (*regional*): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW),

[Continued on next page]

(54) Title: DIAGNOSIS AND PROGNOSIS OF BREAST CANCER PATIENTS



(57) Abstract: The present invention relates to genetic markers whose expression is correlated with breast cancer. Specifically, the invention provides sets of markers whose expression patterns can be used to differentiate clinical conditions associated with breast cancer, such as the presence or absence of the estrogen receptor *ESR1*, and *BRCA1* and sporadic tumors, and to provide information on the likelihood of tumor distant metastases within five years of initial diagnosis. The invention relates to methods of using these markers to distinguish these conditions. The invention also relates to kits containing ready-to-use microarrays and computer software for data analysis using the statistical methods disclosed herein.

WO 02/103320 A2



Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM),
European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR,
GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent
(BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR,
NE, SN, TD, TG).

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

Published:

- *without international search report and to be republished upon receipt of that report*

DIAGNOSIS AND PROGNOSIS OF BREAST CANCER PATIENTS

5

This application claims benefit of United States Provisional Application No. 60/298,918, filed June 18, 2001, and United States Provisional Application No. 60/380,710, filed on May 14, 2002, each of which is incorporated by reference herein in its entirety.

This application includes a Sequence Listing submitted on compact disc,
10 recorded on two compact discs, including one duplicate, containing Filename
9301175228.txt, of size 6,755,971 bytes, created June 13, 2002. The sequence listing on the
compact discs is incorporated by reference herein in its entirety.

1. FIELD OF THE INVENTION

15 The present invention relates to the identification of marker genes useful in
the diagnosis and prognosis of breast cancer. More particularly, the invention relates to the
identification of a set of marker genes associated with breast cancer, a set of marker genes
differentially expressed in estrogen receptor (+) versus estrogen receptor (-) tumors, a set of
marker genes differentially expressed in *BRCA1* versus sporadic tumors, and a set of marker
20 genes differentially expressed in sporadic tumors from patients with good clinical prognosis
(*i.e.*, metastasis- or disease-free >5 years) versus patients with poor clinical prognosis (*i.e.*,
metastasis- or disease-free <5 years). For each of the marker sets above, the invention
further relates to methods of distinguishing the breast cancer-related conditions. The
invention further provides methods for determining the course of treatment of a patient with
25 breast cancer.

2. BACKGROUND OF THE INVENTION

The increased number of cancer cases reported in the United States, and,
indeed, around the world, is a major concern. Currently there are only a handful of
30 treatments available for specific types of cancer, and these provide no guarantee of success.
In order to be most effective, these treatments require not only an early detection of the
malignancy, but a reliable assessment of the severity of the malignancy.

The incidence of breast cancer, a leading cause of death in women, has been
gradually increasing in the United States over the last thirty years. Its cumulative risk is
35 relatively high; 1 in 8 women are expected to develop some type of breast cancer by age 85

in the United States. In fact, breast cancer is the most common cancer in women and the second most common cause of cancer death in the United States. In 1997, it was estimated that 181,000 new cases were reported in the U.S., and that 44,000 people would die of breast cancer (Parker *et al.*, *CA Cancer J. Clin.* 47:5-27 (1997); Chu *et al.*, *J. Nat. Cancer*
5 *Inst.* 88:1571-1579 (1996)). While mechanism of tumorigenesis for most breast carcinomas is largely unknown, there are genetic factors that can predispose some women to developing breast cancer (Miki *et al.*, *Science*, 266:66-71(1994)). The discovery and characterization of *BRCA1* and *BRCA2* has recently expanded our knowledge of genetic factors which can contribute to familial breast cancer. Germ-line mutations within these two loci are
10 associated with a 50 to 85% lifetime risk of breast and/or ovarian cancer (Casey, *Curr. Opin. Oncol.* 9:88-93 (1997); Marcus *et al.*, *Cancer* 77:697-709 (1996)). Only about 5% to 10% of breast cancers are associated with breast cancer susceptibility genes, *BRCA1* and *BRCA2*. The cumulative lifetime risk of breast cancer for women who carry the mutant *BRCA1* is predicted to be approximately 92%, while the cumulative lifetime risk for the
15 non-carrier majority is estimated to be approximately 10%. *BRCA1* is a tumor suppressor gene that is involved in DNA repair and cell cycle control, which are both important for the maintenance of genomic stability. More than 90% of all mutations reported so far result in a premature truncation of the protein product with abnormal or abolished function. The histology of breast cancer in *BRCA1* mutation carriers differs from that in sporadic cases,
20 but mutation analysis is the only way to find the carrier. Like *BRCA1*, *BRCA2* is involved in the development of breast cancer, and like *BRCA1* plays a role in DNA repair. However, unlike *BRCA1*, it is not involved in ovarian cancer.

Other genes have been linked to breast cancer, for example c-erb-2 (*HER2*) and p53 (Beenken *et al.*, *Ann. Surg.* 233(5):630-638 (2001). Overexpression of c-erb-2
25 (*HER2*) and p53 have been correlated with poor prognosis (Rudolph *et al.*, *Hum. Pathol.* 32(3):311-319 (2001), as has been aberrant expression products of *mdm2* (Lukas *et al.*, *Cancer Res.* 61(7):3212-3219 (2001) and cyclin1 and p27 (Porter & Roberts, International Publication WO98/33450, published August 6, 1998). However, no other clinically useful markers consistently associated with breast cancer have been identified.

30 Sporadic tumors, those not currently associated with a known germline mutation, constitute the majority of breast cancers. It is also likely that other, non-genetic factors also have a significant effect on the etiology of the disease. Regardless of the cancer's origin, breast cancer morbidity and mortality increases significantly if it is not detected early in its progression. Thus, considerable effort has focused on the early detection
35 of cellular transformation and tumor formation in breast tissue.

A marker-based approach to tumor identification and characterization promises improved diagnostic and prognostic reliability. Typically, the diagnosis of breast cancer requires histopathological proof of the presence of the tumor. In addition to diagnosis, histopathological examinations also provide information about prognosis and selection of treatment regimens. Prognosis may also be established based upon clinical parameters such as tumor size, tumor grade, the age of the patient, and lymph node metastasis.

Diagnosis and/or prognosis may be determined to varying degrees of effectiveness by direct examination of the outside of the breast, or through mammography or other X-ray imaging methods (Jatoi, *Am. J. Surg.* 177:518-524 (1999)). The latter approach is not without considerable cost, however. Every time a mammogram is taken, the patient incurs a small risk of having a breast tumor induced by the ionizing properties of the radiation used during the test. In addition, the process is expensive and the subjective interpretations of a technician can lead to imprecision. For example, one study showed major clinical disagreements for about one-third of a set of mammograms that were interpreted individually by a surveyed group of radiologists. Moreover, many women find that undergoing a mammogram is a painful experience. Accordingly, the National Cancer Institute has not recommended mammograms for women under fifty years of age, since this group is not as likely to develop breast cancers as are older women. It is compelling to note, however, that while only about 22% of breast cancers occur in women under fifty, data suggests that breast cancer is more aggressive in pre-menopausal women.

In clinical practice, accurate diagnosis of various subtypes of breast cancer is important because treatment options, prognosis, and the likelihood of therapeutic response all vary broadly depending on the diagnosis. Accurate prognosis, or determination of distant metastasis-free survival could allow the oncologist to tailor the administration of adjuvant chemotherapy, with women having poorer prognoses being given the most aggressive treatment. Furthermore, accurate prediction of poor prognosis would greatly impact clinical trials for new breast cancer therapies, because potential study patients could then be stratified according to prognosis. Trials could then be limited to patients having poor prognosis, in turn making it easier to discern if an experimental therapy is efficacious.

To date, no set of satisfactory predictors for prognosis based on the clinical information alone has been identified. The detection of *BRCA1* or *BRCA2* mutations represents a step towards the design of therapies to better control and prevent the appearance of these tumors. However, there is no equivalent means for the diagnosis of

patients with sporadic tumors, the most common type of breast cancer tumor, nor is there a means of differentiating subtypes of breast cancer.

3. SUMMARY OF THE INVENTION

5 The invention provides gene marker sets that distinguish various types and subtypes of breast cancer, and methods of use therefor. In one embodiment, the invention provides a method for classifying a cell sample as ER(+) or ER(-) comprising detecting a difference in the expression of a first plurality of genes relative to a control, said first plurality of genes consisting of at least 5 of the genes corresponding to the markers listed in
10 Table 1. In specific embodiments, said plurality of genes consists of at least 50, 100, 200, 500, 1000, up to 2,460 of the gene markers listed in Table 1. In another specific embodiment, said plurality of genes consists of each of the genes corresponding to the 2,460 markers listed in Table 2. In another specific embodiment, said plurality consists of the 550 markers listed in Table 2. In another specific embodiment, said control comprises nucleic
15 acids derived from a pool of tumors from individual sporadic patients. In another specific embodiment, said detecting comprises the steps of: (a) generating an ER(+) template by hybridization of nucleic acids derived from a plurality of ER(+) patients within a plurality of sporadic patients against nucleic acids derived from a pool of tumors from individual sporadic patients; (b) generating an ER(-) template by hybridization of nucleic acids derived
20 from a plurality of ER(-) patients within said plurality of sporadic patients against nucleic acids derived from said pool of tumors from individual sporadic patients within said plurality; (c) hybridizing nucleic acids derived from an individual sample against said pool; and (d) determining the similarity of marker gene expression in the individual sample to the ER(+) template and the ER(-) template, wherein if said expression is more similar to the
25 ER(+) template, the sample is classified as ER(+), and if said expression is more similar to the ER(-) template, the sample is classified as ER(-).

 The invention further provides the above methods, applied to the classification of samples as *BRCA1* or sporadic, and classifying patients as having good prognosis or poor prognosis. For the *BRCA1*/sporadic gene markers, the invention provides
30 that the method may be used wherein the plurality of genes is at least 5, 20, 50, 100, 200 or 300 of the *BRCA1*/sporadic markers listed in Table 3. In a specific embodiment, the optimum 100 markers listed in Table 4 are used. For the prognostic markers, the invention provides that at least 5, 20, 50, 100, or 200 gene markers listed in Table 5 may be used. In a specific embodiment, the optimum 70 markers listed in Table 6 are used.

35

The invention further provides that markers may be combined. Thus, in one embodiment, at least 5 markers from Table 1 are used in conjunction with at least 5 markers from Table 3. In another embodiment, at least 5 markers from Table 5 are used in conjunction with at least 5 markers from Table 3. In another embodiment, at least 5 markers from Table 1 are used in conjunction with at least 5 markers from Table 5. In another embodiment, at least 5 markers from each of Tables 1, 3, and 5 are used simultaneously.

The invention further provides a method for classifying a sample as ER(+) or ER(-) by calculating the similarity between the expression of at least 5 of the markers listed in Table 1 in the sample to the expression of the same markers in an ER(-) nucleic acid pool and an ER(+) nucleic acid pool, comprising the steps of: (a) labeling nucleic acids derived from a sample, with a first fluorophore to obtain a first pool of fluorophore-labeled nucleic acids; (b) labeling with a second fluorophore a first pool of nucleic acids derived from two or more ER(+) samples, and a second pool of nucleic acids derived from two or more ER(-) samples; (c) contacting said first fluorophore-labeled nucleic acid and said first pool of second fluorophore-labeled nucleic acid with said first microarray under conditions such that hybridization can occur, and contacting said first fluorophore-labeled nucleic acid and said second pool of second fluorophore-labeled nucleic acid with said second microarray under conditions such that hybridization can occur, detecting at each of a plurality of discrete loci on the first microarray a first fluorescent emission signal from said first fluorophore-labeled nucleic acid and a second fluorescent emission signal from said first pool of second fluorophore-labeled genetic matter that is bound to said first microarray under said conditions, and detecting at each of the marker loci on said second microarray said first fluorescent emission signal from said first fluorophore-labeled nucleic acid and a third fluorescent emission signal from said second pool of second fluorophore-labeled nucleic acid; (d) determining the similarity of the sample to the ER(-) and ER(+) pools by comparing said first fluorescence emission signals and said second fluorescence emission signals, and said first emission signals and said third fluorescence emission signals; and (e) classifying the sample as ER(+) where the first fluorescence emission signals are more similar to said second fluorescence emission signals than to said third fluorescent emission signals, and classifying the sample as ER(-) where the first fluorescence emission signals are more similar to said third fluorescence emission signals than to said second fluorescent emission signals, wherein said similarity is defined by a statistical method. The invention further provides that the other disclosed marker sets may be used in the above method to

35

distinguish *BRCA1* from sporadic tumors, and patients with poor prognosis from patients with good prognosis.

In a specific embodiment, said similarity is calculated by determining a first sum of the differences of expression levels for each marker between said first fluorophore-labeled nucleic acid and said first pool of second fluorophore-labeled nucleic acid, and a second sum of the differences of expression levels for each marker between said first fluorophore-labeled nucleic acid and said second pool of second fluorophore-labeled nucleic acid, wherein if said first sum is greater than said second sum, the sample is classified as ER(-), and if said second sum is greater than said first sum, the sample is classified as ER(+). In another specific embodiment, said similarity is calculated by computing a first classifier parameter P_1 between an ER(+) template and the expression of said markers in said sample, and a second classifier parameter P_2 between an ER(-) template and the expression of said markers in said sample, wherein said P_1 and P_2 are calculated according to the formula:

$$P_i = (\vec{z}_i \cdot \vec{y}) / (\|\vec{z}_i\| \cdot \|\vec{y}\|), \quad \text{Equation (1)}$$

wherein \vec{z}_1 and \vec{z}_2 are ER(-) and ER(+) templates, respectively, and are calculated by averaging said second fluorescence emission signal for each of said markers in said first pool of second fluorophore-labeled nucleic acid and said third fluorescence emission signal for each of said markers in said second pool of second fluorophore-labeled nucleic acid, respectively, and wherein \vec{y} is said first fluorescence emission signal of each of said markers in the sample to be classified as ER(+) or ER(-), wherein the expression of the markers in the sample is similar to ER(+) if $P_1 < P_2$, and similar to ER(-) if $P_1 > P_2$.

The invention further provides a method for identifying marker genes the expression of which is associated with a particular phenotype. In one embodiment, the invention provides a method for determining a set of marker genes whose expression is associated with a particular phenotype, comprising the steps of: (a) selecting the phenotype having two or more phenotype categories; (b) identifying a plurality of genes wherein the expression of said genes is correlated or anticorrelated with one of the phenotype categories, and wherein the correlation coefficient for each gene is calculated according to the equation

$$\rho = (\vec{c} \cdot \vec{r}) / (\|\vec{c}\| \cdot \|\vec{r}\|) \quad \text{Equation (2)}$$

wherein \vec{c} is a number representing said phenotype category and \vec{r} is the logarithmic expression ratio across all the samples for each individual gene, wherein if the correlation

coefficient has an absolute value of a threshold value or greater, said expression of said gene is associated with the phenotype category, and wherein said plurality of genes is a set of marker genes whose expression is associated with a particular phenotype. The threshold depends upon the number of samples used; the threshold can be calculated as $3 \times 1/\sqrt{n-3}$, where $1/\sqrt{n-3}$ is the distribution width and n = the number of samples. In a specific embodiment where $n = 98$, said threshold value is 0.3. In a specific embodiment, said set of marker genes is validated by: (a) using a statistical method to randomize the association between said marker genes and said phenotype category, thereby creating a control correlation coefficient for each marker gene; (b) repeating step (a) one hundred or more times to develop a frequency distribution of said control correlation coefficients for each marker gene; (c) determining the number of marker genes having a control correlation coefficient of a threshold value or above, thereby creating a control marker gene set; and (d) comparing the number of control marker genes so identified to the number of marker genes, wherein if the p value of the difference between the number of marker genes and the number of control genes is less than 0.01, said set of marker genes is validated. In another specific embodiment, said set of marker genes is optimized by the method comprising: (a) rank-ordering the genes by amplitude of correlation or by significance of the correlation coefficients, and (b) selecting an arbitrary number of marker genes from the top of the rank-ordered list. The threshold value depends upon the number of samples tested.

The invention further provides a method for assigning a person to one of a plurality of categories in a clinical trial, comprising determining for each said person the level of expression of at least five of the prognosis markers listed in Table 6, determining therefrom whether the person has an expression pattern that correlates with a good prognosis or a poor prognosis, and assigning said person to one category in a clinical trial if said person is determined to have a good prognosis, and a different category if that person is determined to have a poor prognosis. The invention further provides a method for assigning a person to one of a plurality of categories in a clinical trial, where each of said categories is associated with a different phenotype, comprising determining for each said person the level of expression of at least five markers from a set of markers, wherein said set of markers includes markers associated with each of said clinical categories, determining therefrom whether the person has an expression pattern that correlates with one of the clinical categories, an assigning said person to one of said categories if said person is determined to have a phenotype associated with that category.

The invention further provides a method of classifying a first cell or organism as having one of at least two different phenotypes, said at least two different

phenotypes comprising a first phenotype and a second phenotype, said method comprising:
(a) comparing the level of expression of each of a plurality of genes in a first sample from
the first cell or organism to the level of expression of each of said genes, respectively, in a
pooled sample from a plurality of cells or organisms, said plurality of cells or organisms
5 comprising different cells or organisms exhibiting said at least two different phenotypes,
respectively, to produce a first compared value; (b) comparing said first compared value to a
second compared value, wherein said second compared value is the product of a method
comprising comparing the level of expression of each of said genes in a sample from a cell
or organism characterized as having said first phenotype to the level of expression of each
10 of said genes, respectively, in said pooled sample; (c) comparing said first compared value
to a third compared value, wherein said third compared value is the product of a method
comprising comparing the level of expression of each of said genes in a sample from a cell
or organism characterized as having said second phenotype to the level of expression of
each of said genes, respectively, in said pooled sample, (d) optionally carrying out one or
15 more times a step of comparing said first compared value to one or more additional
compared values, respectively, each additional compared value being the product of a
method comprising comparing the level of expression of each of said genes in a sample
from a cell or organism characterized as having a phenotype different from said first and
second phenotypes but included among said at least two different phenotypes, to the level of
20 expression of each of said genes, respectively, in said pooled sample; and (e) determining to
which of said second, third and, if present, one or more additional compared values, said
first compared value is most similar, wherein said first cell or organism is determined to
have the phenotype of the cell or organism used to produce said compared value most
similar to said first compared value.

25 In a specific embodiment of the above method, said compared values are
each ratios of the levels of expression of each of said genes. In another specific
embodiment, each of said levels of expression of each of said genes in said pooled sample
are normalized prior to any of said comparing steps. In another specific embodiment,
normalizing said levels of expression is carried out by dividing each of said levels of
30 expression by the median or mean level of expression of each of said genes or dividing by
the mean or median level of expression of one or more housekeeping genes in said pooled
sample. In a more specific embodiment, said normalized levels of expression are subjected
to a log transform and said comparing steps comprise subtracting said log transform from
the log of said levels of expression of each of said genes in said sample from said cell or
35 organism. In another specific embodiment, said at least two different phenotypes are

different stages of a disease or disorder. In another specific embodiment, said at least two different phenotypes are different prognoses of a disease or disorder. In yet another specific embodiment, said levels of expression of each of said genes, respectively, in said pooled sample or said levels of expression of each of said genes in a sample from said cell or
5 organism characterized as having said first phenotype, said second phenotype, or said phenotype different from said first and second phenotypes, respectively, are stored on a computer.

The invention further provides microarrays comprising the disclosed marker sets. In one embodiment, the invention provides a microarray comprising at least 5 markers
10 derived from any one of Tables 1-6, wherein at least 50% of the probes on the microarray are present in any one of Tables 1-6. In more specific embodiments, at least 60%, 70%, 80%, 90%, 95% or 98% of the probes on said microarray are present in any one of Tables 1-6.

In another embodiment, the invention provides a microarray for
15 distinguishing ER(+) and ER(-) cell samples comprising a positionally-addressable array of polynucleotide probes bound to a support, said polynucleotide probes comprising a plurality of polynucleotide probes of different nucleotide sequences, each of said different nucleotide sequences comprising a sequence complementary and hybridizable to a plurality of genes, said plurality consisting of at least 5 of the genes corresponding to the markers listed in
20 Table 1 or Table 2, wherein at least 50% of the probes on the microarray are present in any one of Table 1 or Table 2. In yet another embodiment, the invention provides a microarray for distinguishing *BRCA1*-type and sporadic tumor-type cell samples comprising a positionally-addressable array of polynucleotide probes bound to a support, said polynucleotide probes comprising a plurality of polynucleotide probes of different
25 nucleotide sequences, each of said different nucleotide sequences comprising a sequence complementary and hybridizable to a plurality of genes, said plurality consisting of at least 5 of the genes corresponding to the markers listed in Table 3 or Table 4, wherein at least 50% of the probes on the microarray are present in any one of Table 3 or Table 4. In still another embodiment, the invention provides a microarray for distinguishing cell samples from
30 patients having a good prognosis and cell samples from patients having a poor prognosis comprising a positionally-addressable array of polynucleotide probes bound to a support, said polynucleotide probes comprising a plurality of polynucleotide probes of different nucleotide sequences, each of said different nucleotide sequences comprising a sequence complementary and hybridizable to a plurality of genes, said plurality consisting of at least 5
35 of the genes corresponding to the markers listed in Table 5 or Table 6, wherein at least 50%

of the probes on the microarray are present in any one of Table 5 or Table 6. The invention further provides for microarrays comprising at least 5, 20, 50, 100, 200, 500, 100, 1,250, 1,500, 1,750, or 2,000 of the ER-status marker genes listed in Table 1, at least 5, 20, 50, 100, 200, or 300 of the *BRCA1* sporadic marker genes listed in Table 3, or at least 5, 20, 50, 100 or 200 of the prognostic marker genes listed in Table 5, in any combination, wherein at least 50%, 60%, 70%, 80%, 90%, 95% or 98% of the probes on said microarrays are present in Table 1, Table 3 and/or Table 5.

The invention further provides a kit for determining the ER-status of a sample, comprising at least two microarrays each comprising at least 5 of the markers listed in Table 1, and a computer system for determining the similarity of the level of nucleic acid derived from the markers listed in Table 1 in a sample to that in an ER(-) pool and an ER(+) pool, the computer system comprising a processor, and a memory encoding one or more programs coupled to the processor, wherein the one or more programs cause the processor to perform a method comprising computing the aggregate differences in expression of each marker between the sample and ER(-) pool and the aggregate differences in expression of each marker between the sample and ER(+) pool, or a method comprising determining the correlation of expression of the markers in the sample to the expression in the ER(-) and ER(+) pools, said correlation calculated according to Equation (4). The invention provides for kits able to distinguish *BRCA1* and sporadic tumors, and samples from patients with good prognosis from samples from patients with poor prognosis, by inclusion of the appropriate marker gene sets. The invention further provides a kit for determining whether a sample is derived from a patient having a good prognosis or a poor prognosis, comprising at least one microarray comprising probes to at least 5 of the genes corresponding to the markers listed in Table 5, and a computer readable medium having recorded thereon one or more programs for determining the similarity of the level of nucleic acid derived from the markers listed in Table 5 in a sample to that in a pool of samples derived from individuals having a good prognosis and a pool of samples derived from individuals having a good prognosis, wherein the one or more programs cause a computer to perform a method comprising computing the aggregate differences in expression of each marker between the sample and the good prognosis pool and the aggregate differences in expression of each marker between the sample and the poor prognosis pool, or a method comprising determining the correlation of expression of the markers in the sample to the expression in the good prognosis and poor prognosis pools, said correlation calculated according to Equation (3).

35

4. BRIEF DESCRIPTION OF THE FIGURES

FIG. 1 is a Venn-type diagram showing the overlap between the marker sets disclosed herein, including the 2,460 ER markers, the 430 *BRCA1*/sporadic markers, and the 231 prognosis reporters.

FIG. 2 shows the experimental procedures for measuring differential changes in mRNA transcript abundance in breast cancer tumors used in this study. In each experiment, Cy5-labeled cRNA from one tumor X is hybridized on a 25k human microarray together with a Cy3-labeled cRNA pool made of cRNA samples from tumors 1, 2, ... N. The digital expression data were obtained by scanning and image processing. The error modeling allowed us to assign a p-value to each transcript ratio measurement.

FIG. 3 Two-dimensional clustering reveals two distinctive types of tumors. The clustering was based on the gene expression data of 98 breast cancer tumors over 4986 significant genes. Dark gray (red) presents up-regulation, light gray (green) represents down-regulation, black indicates no change in expression, and gray indicates that data is not available. 4986 genes were selected that showed a more than two fold change in expression ratios in more than five experiments. Selected clinical data for test results of BR CA1 mutations, estrogen receptor (ER), and proestrogen receptor (PR), tumor grade, lymphocytic infiltrate, and angiogenesis are shown at right. Black denotes negative and white denotes positive. The dominant pattern in the lower part consists of 36 patients, out of which 34 are ER-negative (total 39), and 16 are BR CA1-mutation carriers (total 18).

FIG. 4 A portion of unsupervised clustered results as shown in FIG. 3. *ESR1* (the estrogen receptor gene) is coregulated with a set of genes that are strongly co-regulated to form a dominant pattern.

FIG. 5A Histogram of correlation coefficients of significant genes between their expression ratios and estrogen-receptor (ER) status (*i.e.*, ER level). The histogram for experimental data is shown as a gray line. The results of one Monte-Carlo trial is shown in solid black. There are 2,460 genes whose expression data correlate with ER status at a level higher than 0.3 or anti-correlated with ER status at a level lower than -0.3.

FIG. 5B The distribution of the number of genes that satisfied the same selection criteria (amplitude of correlation above 0.3) from 10,000 Monte-Carlo runs. It is estimated that this set of 2,460 genes reports ER status at a confidence level of $p > 99.99\%$.

FIG. 6 Classification Type 1 and Type 2 error rates as a function of the number (out of 2,460) marker genes used in the classifier. The combined error rate is lowest when approximately 550 marker genes are used.

FIG. 7 Classification of 98 tumor samples as ER(+) or ER(-) based on expression levels of the 550 optimal marker genes. ER(+) samples (above white line) exhibit a clearly different expression pattern than ER(-) samples (below white line).

FIG. 8 Correlation between expression levels in samples from each patient and the average profile of the ER(-) group vs. correlation with the ER(+) group. Squares represent samples from clinically ER(-) patients; dots represent samples from clinically ER(+) patients.

FIG. 9A Histogram of correlation coefficients of gene expression ratio of each significant gene with the *BRCA1* mutation status is shown as a solid line. The dashed line indicates a frequency distribution obtained from one Monte-Carlo run. 430 genes exhibited an amplitude of correlation or anti-correlation greater than 0.35.

FIG. 9B Frequency distribution of the number of genes that exhibit an amplitude of correlation or anti-correlation greater than 0.35 for the 10,000 Monte-Carlo run control. Mean = 115. $p(n > 430) = 0.48\%$ and $p(>430/2) = 9.0\%$.

FIG. 10 Classification type 1 and type 2 error rates as a function of the number of discriminating genes used in the classifier (template). The combined error rate is lowest when approximately 100 discriminating marker genes are used.

FIG. 11A The classification of 38 tumors in the ER(-) group into two subgroups, *BRCA1* and sporadic, by using the optimal set of 100 discriminating marker genes. Patients above the white line are characterized by *BRCA1*-related patterns.

FIG. 11B Correlation between expression levels in samples from each ER(-) patient and the average profile of the *BRCA1* group vs. correlation with the sporadic group. Squares represent samples from patients with sporadic-type tumors; dots represent samples from patients carrying the *BRCA1* mutation.

FIG. 12A Histogram of correlation coefficients of gene expression ratio of each significant gene with the prognostic category (distant metastases group and no distant metastases group) is shown as a solid line. The distribution obtained from one Monte-Carlo run is shown as a dashed line. The amplitude of correlation or anti-correlation of 231 marker genes is greater than 0.3.

FIG. 12B Frequency distribution of the number of genes whose amplitude of correlation or anti-correlation was greater than 0.3 for 10,000 Monte-Carlo runs.

FIG. 13 The distant metastases group classification error rate for type 1 and type 2 as a function of the number of discriminating genes used in the classifier. The combined error rate is lowest when approximately 70 discriminating marker genes are used.

FIG. 14 Classification of 78 sporadic tumors into two prognostic groups, distant metastases (poor prognosis) and no distant metastases (good prognosis) using the optimal set of 70 discriminating marker genes. Patients above the white line are characterized by good prognosis. Patients below the white line are characterized by poor prognosis.

FIG. 15 Correlation between expression levels in samples from each patient and the average profile of the good prognosis group vs. correlation with the poor prognosis group. Squares represent samples from patients having a poor prognosis; dots represent samples from patients having a good prognosis. Red squares represent the 'reoccurred' patients and the blue dots represent the 'non-reoccurred'. A total of 13 out of 78 were misclassified.

FIG. 16 The reoccurrence probability as a function of time since diagnosis. Group A and group B were predicted by using a leave-one-out method based on the optimal set of 70 discriminating marker genes. The 43 patients in group A consists of 37 patients from the no distant metastases group and 6 patients from the distant metastases group. The 35 patients in group B consists of 28 patients from the distant metastases group and 7 patients from the no distant metastases group.

FIG. 17 The distant metastases probability as a function of time since diagnosis for ER(+) (yes) or ER(-) (no) individuals.

FIG. 18 The distant metastases probability as a function of time since diagnosis for progesterone receptor (PR)(+) (yes) or PR(-) (no) individuals.

FIG. 19A, B The distant metastases probability as a function of time since diagnosis. Groups were defined by the tumor grades.

FIG. 20A Classification of 19 independent sporadic tumors into two prognostic groups, distant metastases and no distant metastases, using the 70 optimal marker genes. Patients above the white line have a good prognosis. Patients below the white line have a poor prognosis.

FIG. 20B Correlation between expression ratios of each patient and the average expression ratio of the good prognosis group is defined by the training set versus the correlation between expression ratios of each patient and the average expression ratio of the poor prognosis training set. Of nine patients in the good prognosis group, three are from the "distant metastases group"; of ten patients in the good prognosis group, one patient is from the "no distant metastases group". This error rate of 4 out of 19 is consistent with 13 out of 78 for the initial 78 patients.

FIG. 20C The reoccurrence probability as a function of time since diagnosis for two groups predicted based on expression of the optimal 70 marker genes.

FIG. 21A Sensitivity vs. 1-specificity for good prognosis classification.

FIG. 21B Sensitivity vs. 1-specificity for poor prognosis classification.

5 FIG. 21C Total error rate as a function of threshold on the modeled likelihood. Six clinical parameters (ER status, PR status, tumor grade, tumor size, patient age, and presence or absence of angioinvasion) were used to perform the clinical modeling.

FIG. 22 Comparison of the log(ratio) of individual samples using the “material sample pool” vs. mean subtracted log(intensity) using the “mathematical sample pool” for 70 reporter genes in the 78 sporadic tumor samples. The “material sample pool” was constructed from the 78 sporadic tumor samples.

FIG. 23A Results of the “leave one out” cross validation based on single channel data. Samples are grouped according to each sample’s coefficient of correlation to the average “good prognosis” profile and “poor prognosis” profile for the 70 genes examined. The white line separates samples from patients classified as having poor prognoses (below) and good prognoses (above).

FIG. 23B Scatter plot of coefficients of correlation to the average expression in “good prognosis” samples and “poor prognosis” samples. The false positive rate (*i.e.*, rate of incorrectly classifying a sample as being from a patient having a good prognosis as being one from a patient having a poor prognosis) was 10 out of 44, and the false negative rate is 6 out of 34.

FIG. 24A Single-channel hybridization data for samples ranked according to the coefficients of correlation with the good prognosis classifier. Samples classified as “good prognosis” lie above the white line, and those classified as “poor prognosis” lie below.

FIG. 24B Scatterplot of sample correlation coefficients, with three incorrectly classified samples lying to the right of the threshold correlation coefficient value. The threshold correlation value was set at 0.2727 to limit the false negatives to approximately 10% of the samples.

30

5. DETAILED DESCRIPTION OF THE INVENTION

5.1 INTRODUCTION

The invention relates to sets of genetic markers whose expression patterns correlate with important characteristics of breast cancer tumors. *i.e.*, estrogen receptor (ER) status, *BRCA1* status, and the likelihood of relapse (*i.e.*, distant metastasis or poor

35

prognosis). More specifically, the invention provides for sets of genetic markers that can distinguish the following three clinical conditions. First, the invention relates to sets of markers whose expression correlates with the ER status of a patient, and which can be used to distinguish ER(+) from ER(-) patients. ER status is a useful prognostic indicator, and an indicator of the likelihood that a patient will respond to certain therapies, such as tamoxifen. Also, among women who are ER positive the response rate (over 50%) to hormonal therapy is much higher than the response rate (less 10%) in patients whose ER status is negative. In patients with ER positive tumors the possibility of achieving a hormonal response is directly proportional to the level ER (P. Clabresi and P.S. Schein, MEDICAL ONCOLOGY (2ND ED.), McGraw-Hill, Inc., New York (1993)). Second, the invention further relates to sets of markers whose expression correlates with the presence of *BRCA1* mutations, and which can be used to distinguish *BRCA1*-type tumors from sporadic tumors. Third, the invention relates to genetic markers whose expression correlates with clinical prognosis, and which can be used to distinguish patients having good prognoses (*i.e.*, no distant metastases of a tumor within five years) from poor prognoses (*i.e.*, distant metastases of a tumor within five years). Methods are provided for use of these markers to distinguish between these patient groups, and to determine general courses of treatment. Microarrays comprising these markers are also provided, as well as methods of constructing such microarrays. Each markers correspond to a gene in the human genome, *i.e.*, such marker is identifiable as all or a portion of a gene. Finally, because each of the above markers correlates with a certain breast cancer-related conditions, the markers, or the proteins they encode, are likely to be targets for drugs against breast cancer.

5.2 DEFINITIONS

As used herein, "*BRCA1* tumor" means a tumor having cells containing a mutation of the *BRCA1* locus.

The "absolute amplitude" of correlation expressions means the distance, either positive or negative, from a zero value; *i.e.*, both correlation coefficients -0.35 and 0.35 have an absolute amplitude of 0.35.

"Status" means a state of gene expression of a set of genetic markers whose expression is strongly correlated with a particular phenotype. For example, "ER status" means a state of gene expression of a set of genetic markers whose expression is strongly correlated with that of *ESR1* (estrogen receptor gene), wherein the pattern of these genes' expression differs detectably between tumors expressing the receptor and tumors not expressing the receptor.

“Good prognosis” means that a patient is expected to have no distant metastases of a breast tumor within five years of initial diagnosis of breast cancer.

“Poor prognosis” means that a patient is expected to have distant metastases of a breast tumor within five years of initial diagnosis of breast cancer.

5 “Marker” means an entire gene, or an EST derived from that gene, the expression or level of which changes between certain conditions. Where the expression of the gene correlates with a certain condition, the gene is a marker for that condition.

“Marker-derived polynucleotides” means the RNA transcribed from a marker gene, any cDNA or cRNA produced therefrom, and any nucleic acid derived therefrom,
10 such as synthetic nucleic acid having a sequence derived from the gene corresponding to the marker gene.

5.3 MARKERS USEFUL IN DIAGNOSIS AND PROGNOSIS OF BREAST CANCER

5.3.1 MARKER SETS

15 The invention provides a set of 4,986 genetic markers whose expression is correlated with the existence of breast cancer by clustering analysis. A subset of these markers identified as useful for diagnosis or prognosis is listed as SEQ ID NOS: 1-2,699. The invention also provides a method of using these markers to distinguish tumor types in diagnosis or prognosis.

20 In one embodiment, the invention provides a set of 2,460 genetic markers that can classify breast cancer patients by estrogen receptor (ER) status; *i.e.*, distinguish between ER(+) and ER(-) patients or tumors derived from these patients. ER status is an important indicator of the likelihood of a patient’s response to some chemotherapies (*i.e.*, tamoxifen). These markers are listed in Table 1. The invention also provides subsets of at
25 least 5, 10, 25, 50, 100, 200, 300, 400, 500, 750, 1,000, 1,250, 1,500, 1,750 or 2,000 genetic markers, drawn from the set of 2,460 markers, which also distinguish ER(+) and ER(-) patients or tumors. Preferably, the number of markers is 550. The invention further provides a set of 550 of the 2,460 markers that are optimal for distinguishing ER status (Table 2). The invention also provides a method of using these markers to distinguish
30 between ER(+) and ER(-) patients or tumors derived therefrom.

In another embodiment, the invention provides a set of 430 genetic markers that can classify ER(-) breast cancer patients by *BRCA1* status; *i.e.*, distinguish between tumors containing a *BRCA1* mutation and sporadic tumors. These markers are listed in Table 3. The invention further provides subsets of at least 5, 10 20, 30, 40, 50, 75, 100,
35 150, 200, 250, 300 or 350 markers, drawn from the set of 430 markers, which also

distinguish between tumors containing a *BRCA1* mutation and sporadic tumors. Preferably, the number of markers is 100. A preferred set of 100 markers is provided in Table 4. The invention also provides a method of using these markers to distinguish between *BRCA1* and sporadic patients or tumors derived therefrom.

5 In another embodiment, the invention provides a set of 231 genetic markers that can distinguish between patients with a good breast cancer prognosis (no breast cancer tumor distant metastases within five years) and patients with a poor breast cancer prognosis (tumor distant metastases within five years). These markers are listed in Table 5. The invention also provides subsets of at least 5, 10, 20, 30, 40, 50, 75, 100, 150 or 200 markers,
10 drawn from the set of 231, which also distinguish between patients with good and poor prognosis. A preferred set of 70 markers is provided in Table 6. In a specific embodiment, the set of markers consists of the twelve kinase-related markers and the seven cell division- or mitosis-related markers listed. The invention also provides a method of using the above markers to distinguish between patients with good or poor prognosis.

15

20

25

30

35

Table 1. 2,460 gene markers that distinguish ER(+) and ER(-) cell samples.

| | GenBank | | GenBank | |
|----|------------------|--------------|------------------|----------------|
| | Accession Number | SEQ ID NO | Accession Number | SEQ ID NO |
| 5 | AA555029_RC | SEQ ID NO 1 | NM_006984 | SEQ ID NO 1344 |
| | AB000509 | SEQ ID NO 2 | NM_007005 | SEQ ID NO 1345 |
| | AB001451 | SEQ ID NO 3 | NM_007006 | SEQ ID NO 1346 |
| | AB002301 | SEQ ID NO 4 | NM_007019 | SEQ ID NO 1347 |
| | AB002308 | SEQ ID NO 5 | NM_007027 | SEQ ID NO 1348 |
| 10 | AB002351 | SEQ ID NO 6 | NM_007044 | SEQ ID NO 1350 |
| | AB002448 | SEQ ID NO 7 | NM_007050 | SEQ ID NO 1351 |
| | AB006628 | SEQ ID NO 9 | NM_007057 | SEQ ID NO 1352 |
| | AB006630 | SEQ ID NO 10 | NM_007069 | SEQ ID NO 1353 |
| | AB006746 | SEQ ID NO 11 | NM_007074 | SEQ ID NO 1355 |
| 15 | AB007458 | SEQ ID NO 12 | NM_007088 | SEQ ID NO 1356 |
| | AB007855 | SEQ ID NO 13 | NM_007111 | SEQ ID NO 1357 |
| | AB007857 | SEQ ID NO 14 | NM_007146 | SEQ ID NO 1358 |
| | AB007863 | SEQ ID NO 15 | NM_007173 | SEQ ID NO 1359 |
| | AB007883 | SEQ ID NO 16 | NM_007177 | SEQ ID NO 1360 |
| 20 | AB007896 | SEQ ID NO 17 | NM_007196 | SEQ ID NO 1361 |
| | AB007899 | SEQ ID NO 18 | NM_007203 | SEQ ID NO 1362 |
| | AB007916 | SEQ ID NO 19 | NM_007214 | SEQ ID NO 1363 |
| | AB007950 | SEQ ID NO 20 | NM_007217 | SEQ ID NO 1364 |
| | AB011087 | SEQ ID NO 21 | NM_007231 | SEQ ID NO 1365 |
| 25 | AB011089 | SEQ ID NO 22 | NM_007268 | SEQ ID NO 1367 |
| | AB011104 | SEQ ID NO 23 | NM_007274 | SEQ ID NO 1368 |
| | AB011105 | SEQ ID NO 24 | NM_007275 | SEQ ID NO 1369 |
| | AB011121 | SEQ ID NO 25 | NM_007281 | SEQ ID NO 1370 |
| | AB011132 | SEQ ID NO 26 | NM_007309 | SEQ ID NO 1371 |
| 30 | AB011152 | SEQ ID NO 27 | NM_007315 | SEQ ID NO 1372 |
| | AB011179 | SEQ ID NO 28 | NM_007334 | SEQ ID NO 1373 |
| | AB014534 | SEQ ID NO 29 | NM_007358 | SEQ ID NO 1374 |
| | AB014568 | SEQ ID NO 30 | NM_009585 | SEQ ID NO 1375 |
| | AB018260 | SEQ ID NO 31 | NM_009587 | SEQ ID NO 1376 |
| 35 | AB018268 | SEQ ID NO 32 | NM_009588 | SEQ ID NO 1377 |
| | AB018289 | SEQ ID NO 33 | NM_012062 | SEQ ID NO 1378 |
| | AB018345 | SEQ ID NO 35 | NM_012067 | SEQ ID NO 1379 |
| | AB020677 | SEQ ID NO 36 | NM_012101 | SEQ ID NO 1380 |
| | AB020689 | SEQ ID NO 37 | NM_012105 | SEQ ID NO 1381 |
| | AB020695 | SEQ ID NO 38 | NM_012108 | SEQ ID NO 1382 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|--------------|-----------------------------|----------------|
| | AB020710 | SEQ ID NO 39 | NM_012110 | SEQ ID NO 1383 |
| | AB023139 | SEQ ID NO 40 | NM_012124 | SEQ ID NO 1384 |
| 5 | AB023151 | SEQ ID NO 41 | NM_012142 | SEQ ID NO 1386 |
| | AB023152 | SEQ ID NO 42 | NM_012155 | SEQ ID NO 1388 |
| | AB023163 | SEQ ID NO 43 | NM_012175 | SEQ ID NO 1389 |
| | AB023173 | SEQ ID NO 44 | NM_012177 | SEQ ID NO 1390 |
| | AB023211 | SEQ ID NO 45 | NM_012205 | SEQ ID NO 1391 |
| | AB024704 | SEQ ID NO 46 | NM_012219 | SEQ ID NO 1393 |
| 10 | AB028985 | SEQ ID NO 47 | NM_012242 | SEQ ID NO 1394 |
| | AB028986 | SEQ ID NO 48 | NM_012250 | SEQ ID NO 1395 |
| | AB028998 | SEQ ID NO 49 | NM_012261 | SEQ ID NO 1397 |
| | AB029031 | SEQ ID NO 51 | NM_012286 | SEQ ID NO 1398 |
| | AB032951 | SEQ ID NO 52 | NM_012319 | SEQ ID NO 1400 |
| 15 | AB032966 | SEQ ID NO 53 | NM_012332 | SEQ ID NO 1401 |
| | AB032969 | SEQ ID NO 54 | NM_012336 | SEQ ID NO 1402 |
| | AB032977 | SEQ ID NO 56 | NM_012339 | SEQ ID NO 1404 |
| | AB033007 | SEQ ID NO 58 | NM_012341 | SEQ ID NO 1405 |
| | AB033034 | SEQ ID NO 59 | NM_012391 | SEQ ID NO 1406 |
| | AB033035 | SEQ ID NO 60 | NM_012394 | SEQ ID NO 1407 |
| 20 | AB033040 | SEQ ID NO 61 | NM_012413 | SEQ ID NO 1408 |
| | AB033049 | SEQ ID NO 63 | NM_012421 | SEQ ID NO 1409 |
| | AB033050 | SEQ ID NO 64 | NM_012425 | SEQ ID NO 1410 |
| | AB033053 | SEQ ID NO 65 | NM_012427 | SEQ ID NO 1411 |
| | AB033055 | SEQ ID NO 66 | NM_012429 | SEQ ID NO 1413 |
| 25 | AB033058 | SEQ ID NO 67 | NM_012446 | SEQ ID NO 1414 |
| | AB033073 | SEQ ID NO 68 | NM_012463 | SEQ ID NO 1415 |
| | AB033092 | SEQ ID NO 69 | NM_012474 | SEQ ID NO 1416 |
| | AB033111 | SEQ ID NO 70 | NM_013230 | SEQ ID NO 1417 |
| | AB036063 | SEQ ID NO 71 | NM_013233 | SEQ ID NO 1418 |
| | AB037720 | SEQ ID NO 72 | NM_013238 | SEQ ID NO 1419 |
| 30 | AB037743 | SEQ ID NO 74 | NM_013239 | SEQ ID NO 1420 |
| | AB037745 | SEQ ID NO 75 | NM_013242 | SEQ ID NO 1421 |
| | AB037756 | SEQ ID NO 76 | NM_013257 | SEQ ID NO 1423 |
| | AB037765 | SEQ ID NO 77 | NM_013261 | SEQ ID NO 1424 |
| | AB037778 | SEQ ID NO 78 | NM_013262 | SEQ ID NO 1425 |
| 35 | AB037791 | SEQ ID NO 79 | NM_013277 | SEQ ID NO 1426 |
| | AB037793 | SEQ ID NO 80 | NM_013296 | SEQ ID NO 1427 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|---------------|-----------------------------|----------------|
| | AB037802 | SEQ ID NO 81 | NM_013301 | SEQ ID NO 1428 |
| | AB037806 | SEQ ID NO 82 | NM_013324 | SEQ ID NO 1429 |
| 5 | AB037809 | SEQ ID NO 83 | NM_013327 | SEQ ID NO 1430 |
| | AB037836 | SEQ ID NO 84 | NM_013336 | SEQ ID NO 1431 |
| | AB037844 | SEQ ID NO 85 | NM_013339 | SEQ ID NO 1432 |
| | AB037845 | SEQ ID NO 86 | NM_013363 | SEQ ID NO 1433 |
| | AB037848 | SEQ ID NO 87 | NM_013378 | SEQ ID NO 1435 |
| | AB037863 | SEQ ID NO 88 | NM_013384 | SEQ ID NO 1436 |
| 10 | AB037864 | SEQ ID NO 89 | NM_013385 | SEQ ID NO 1437 |
| | AB040881 | SEQ ID NO 90 | NM_013406 | SEQ ID NO 1438 |
| | AB040900 | SEQ ID NO 91 | NM_013437 | SEQ ID NO 1439 |
| | AB040914 | SEQ ID NO 92 | NM_013451 | SEQ ID NO 1440 |
| | AB040926 | SEQ ID NO 93 | NM_013943 | SEQ ID NO 1441 |
| 15 | AB040955 | SEQ ID NO 94 | NM_013994 | SEQ ID NO 1442 |
| | AB040961 | SEQ ID NO 95 | NM_013995 | SEQ ID NO 1443 |
| | AF000974 | SEQ ID NO 97 | NM_014026 | SEQ ID NO 1444 |
| | AF005487 | SEQ ID NO 98 | NM_014029 | SEQ ID NO 1445 |
| | AF007153 | SEQ ID NO 99 | NM_014036 | SEQ ID NO 1446 |
| | AF007155 | SEQ ID NO 100 | NM_014062 | SEQ ID NO 1447 |
| 20 | AF015041 | SEQ ID NO 101 | NM_014074 | SEQ ID NO 1448 |
| | AF016004 | SEQ ID NO 102 | NM_014096 | SEQ ID NO 1450 |
| | AF016495 | SEQ ID NO 103 | NM_014109 | SEQ ID NO 1451 |
| | AF020919 | SEQ ID NO 104 | NM_014112 | SEQ ID NO 1452 |
| | AF026941 | SEQ ID NO 105 | NM_014147 | SEQ ID NO 1453 |
| 25 | AF035191 | SEQ ID NO 106 | NM_014149 | SEQ ID NO 1454 |
| | AF035284 | SEQ ID NO 107 | NM_014164 | SEQ ID NO 1455 |
| | AF035318 | SEQ ID NO 108 | NM_014172 | SEQ ID NO 1456 |
| | AF038182 | SEQ ID NO 109 | NM_014175 | SEQ ID NO 1457 |
| | AF038193 | SEQ ID NO 110 | NM_014181 | SEQ ID NO 1458 |
| | AF042838 | SEQ ID NO 111 | NM_014184 | SEQ ID NO 1459 |
| 30 | AF044127 | SEQ ID NO 112 | NM_014211 | SEQ ID NO 1460 |
| | AF045229 | SEQ ID NO 113 | NM_014214 | SEQ ID NO 1461 |
| | AF047002 | SEQ ID NO 114 | NM_014216 | SEQ ID NO 1462 |
| | AF047826 | SEQ ID NO 115 | NM_014241 | SEQ ID NO 1463 |
| | AF049460 | SEQ ID NO 116 | NM_014246 | SEQ ID NO 1465 |
| 35 | AF052101 | SEQ ID NO 117 | NM_014268 | SEQ ID NO 1466 |
| | AF052117 | SEQ ID NO 118 | NM_014272 | SEQ ID NO 1467 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|---------------|-----------------------------|----------------|
| | AF052155 | SEQ ID NO 119 | NM_014274 | SEQ ID NO 1468 |
| | AF052159 | SEQ ID NO 120 | NM_014289 | SEQ ID NO 1469 |
| 5 | AF052176 | SEQ ID NO 122 | NM_014298 | SEQ ID NO 1470 |
| | AF052185 | SEQ ID NO 123 | NM_014302 | SEQ ID NO 1471 |
| | AF055270 | SEQ ID NO 126 | NM_014315 | SEQ ID NO 1473 |
| | AF058075 | SEQ ID NO 127 | NM_014316 | SEQ ID NO 1474 |
| | AF061034 | SEQ ID NO 128 | NM_014317 | SEQ ID NO 1475 |
| | AF063725 | SEQ ID NO 129 | NM_014320 | SEQ ID NO 1476 |
| 10 | AF063936 | SEQ ID NO 130 | NM_014321 | SEQ ID NO 1477 |
| | AF065241 | SEQ ID NO 131 | NM_014325 | SEQ ID NO 1478 |
| | AF067972 | SEQ ID NO 132 | NM_014335 | SEQ ID NO 1479 |
| | AF070536 | SEQ ID NO 133 | NM_014363 | SEQ ID NO 1480 |
| | AF070552 | SEQ ID NO 134 | NM_014364 | SEQ ID NO 1481 |
| 15 | AF070617 | SEQ ID NO 135 | NM_014365 | SEQ ID NO 1482 |
| | AF073770 | SEQ ID NO 138 | NM_014373 | SEQ ID NO 1483 |
| | AF076612 | SEQ ID NO 139 | NM_014382 | SEQ ID NO 1484 |
| | AF079529 | SEQ ID NO 140 | NM_014395 | SEQ ID NO 1485 |
| | AF090913 | SEQ ID NO 142 | NM_014398 | SEQ ID NO 1486 |
| | AF095719 | SEQ ID NO 143 | NM_014399 | SEQ ID NO 1487 |
| 20 | AF098641 | SEQ ID NO 144 | NM_014402 | SEQ ID NO 1488 |
| | AF099032 | SEQ ID NO 145 | NM_014428 | SEQ ID NO 1489 |
| | AF100756 | SEQ ID NO 146 | NM_014448 | SEQ ID NO 1490 |
| | AF101051 | SEQ ID NO 147 | NM_014449 | SEQ ID NO 1491 |
| | AF103375 | SEQ ID NO 148 | NM_014450 | SEQ ID NO 1492 |
| 25 | AF103458 | SEQ ID NO 149 | NM_014452 | SEQ ID NO 1493 |
| | AF103530 | SEQ ID NO 150 | NM_014453 | SEQ ID NO 1494 |
| | AF103804 | SEQ ID NO 151 | NM_014456 | SEQ ID NO 1495 |
| | AF111849 | SEQ ID NO 152 | NM_014479 | SEQ ID NO 1497 |
| | AF112213 | SEQ ID NO 153 | NM_014501 | SEQ ID NO 1498 |
| | AF113132 | SEQ ID NO 154 | NM_014552 | SEQ ID NO 1500 |
| 30 | AF116682 | SEQ ID NO 156 | NM_014553 | SEQ ID NO 1501 |
| | AF118224 | SEQ ID NO 157 | NM_014570 | SEQ ID NO 1502 |
| | AF118274 | SEQ ID NO 158 | NM_014575 | SEQ ID NO 1503 |
| | AF119256 | SEQ ID NO 159 | NM_014585 | SEQ ID NO 1504 |
| | AF119665 | SEQ ID NO 160 | NM_014595 | SEQ ID NO 1505 |
| 35 | AF121255 | SEQ ID NO 161 | NM_014624 | SEQ ID NO 1507 |
| | AF131748 | SEQ ID NO 162 | NM_014633 | SEQ ID NO 1508 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|---------------|-----------------------------|----------------|
| | AF131753 | SEQ ID NO 163 | NM_014640 | SEQ ID NO 1509 |
| | AF131760 | SEQ ID NO 164 | NM_014642 | SEQ ID NO 1510 |
| 5 | AF131784 | SEQ ID NO 165 | NM_014643 | SEQ ID NO 1511 |
| | AF131828 | SEQ ID NO 166 | NM_014656 | SEQ ID NO 1512 |
| | AF135168 | SEQ ID NO 167 | NM_014668 | SEQ ID NO 1513 |
| | AF141882 | SEQ ID NO 168 | NM_014669 | SEQ ID NO 1514 |
| | AF148505 | SEQ ID NO 169 | NM_014673 | SEQ ID NO 1515 |
| | AF149785 | SEQ ID NO 170 | NM_014675 | SEQ ID NO 1516 |
| 10 | AF151810 | SEQ ID NO 171 | NM_014679 | SEQ ID NO 1517 |
| | AF152502 | SEQ ID NO 172 | NM_014680 | SEQ ID NO 1518 |
| | AF155120 | SEQ ID NO 174 | NM_014696 | SEQ ID NO 1519 |
| | AF159092 | SEQ ID NO 175 | NM_014700 | SEQ ID NO 1520 |
| | AF161407 | SEQ ID NO 176 | NM_014715 | SEQ ID NO 1521 |
| 15 | AF161553 | SEQ ID NO 177 | NM_014721 | SEQ ID NO 1522 |
| | AF164104 | SEQ ID NO 178 | NM_014737 | SEQ ID NO 1524 |
| | AF167706 | SEQ ID NO 179 | NM_014738 | SEQ ID NO 1525 |
| | AF175387 | SEQ ID NO 180 | NM_014747 | SEQ ID NO 1526 |
| | AF176012 | SEQ ID NO 181 | NM_014750 | SEQ ID NO 1527 |
| | AF186780 | SEQ ID NO 182 | NM_014754 | SEQ ID NO 1528 |
| 20 | AF217508 | SEQ ID NO 184 | NM_014767 | SEQ ID NO 1529 |
| | AF220492 | SEQ ID NO 185 | NM_014770 | SEQ ID NO 1530 |
| | AF224266 | SEQ ID NO 186 | NM_014773 | SEQ ID NO 1531 |
| | AF230904 | SEQ ID NO 187 | NM_014776 | SEQ ID NO 1532 |
| | AF234532 | SEQ ID NO 188 | NM_014782 | SEQ ID NO 1533 |
| 25 | AF257175 | SEQ ID NO 189 | NM_014785 | SEQ ID NO 1534 |
| | AF257659 | SEQ ID NO 190 | NM_014791 | SEQ ID NO 1535 |
| | AF272357 | SEQ ID NO 191 | NM_014808 | SEQ ID NO 1536 |
| | AF279865 | SEQ ID NO 192 | NM_014811 | SEQ ID NO 1537 |
| | AI497657_RC | SEQ ID NO 193 | NM_014812 | SEQ ID NO 1538 |
| | AJ012755 | SEQ ID NO 194 | NM_014838 | SEQ ID NO 1540 |
| 30 | AJ223353 | SEQ ID NO 195 | NM_014862 | SEQ ID NO 1542 |
| | AJ224741 | SEQ ID NO 196 | NM_014865 | SEQ ID NO 1543 |
| | AJ224864 | SEQ ID NO 197 | NM_014870 | SEQ ID NO 1544 |
| | AJ225092 | SEQ ID NO 198 | NM_014875 | SEQ ID NO 1545 |
| | AJ225093 | SEQ ID NO 199 | NM_014886 | SEQ ID NO 1547 |
| 35 | AJ249377 | SEQ ID NO 200 | NM_014889 | SEQ ID NO 1548 |
| | AJ270996 | SEQ ID NO 202 | NM_014905 | SEQ ID NO 1549 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|---------------|-----------------------------|----------------|
| | AJ272057 | SEQ ID NO 203 | NM_014935 | SEQ ID NO 1550 |
| | AJ275978 | SEQ ID NO 204 | NM_014945 | SEQ ID NO 1551 |
| 5 | AJ276429 | SEQ ID NO 205 | NM_014965 | SEQ ID NO 1552 |
| | AK000004 | SEQ ID NO 206 | NM_014967 | SEQ ID NO 1553 |
| | AK000005 | SEQ ID NO 207 | NM_014968 | SEQ ID NO 1554 |
| | AK000106 | SEQ ID NO 208 | NM_015032 | SEQ ID NO 1555 |
| | AK000142 | SEQ ID NO 209 | NM_015239 | SEQ ID NO 1556 |
| | AK000168 | SEQ ID NO 210 | NM_015383 | SEQ ID NO 1557 |
| 10 | AK000345 | SEQ ID NO 212 | NM_015392 | SEQ ID NO 1558 |
| | AK000543 | SEQ ID NO 213 | NM_015416 | SEQ ID NO 1559 |
| | AK000552 | SEQ ID NO 214 | NM_015417 | SEQ ID NO 1560 |
| | AK000643 | SEQ ID NO 216 | NM_015420 | SEQ ID NO 1561 |
| | AK000660 | SEQ ID NO 217 | NM_015434 | SEQ ID NO 1562 |
| 15 | AK000689 | SEQ ID NO 218 | NM_015474 | SEQ ID NO 1563 |
| | AK000770 | SEQ ID NO 220 | NM_015507 | SEQ ID NO 1565 |
| | AK000933 | SEQ ID NO 221 | NM_015513 | SEQ ID NO 1566 |
| | AK001100 | SEQ ID NO 223 | NM_015515 | SEQ ID NO 1567 |
| | AK001164 | SEQ ID NO 224 | NM_015523 | SEQ ID NO 1568 |
| | AK001166 | SEQ ID NO 225 | NM_015524 | SEQ ID NO 1569 |
| 20 | AK001295 | SEQ ID NO 226 | NM_015599 | SEQ ID NO 1571 |
| | AK001380 | SEQ ID NO 227 | NM_015623 | SEQ ID NO 1572 |
| | AK001423 | SEQ ID NO 228 | NM_015640 | SEQ ID NO 1573 |
| | AK001438 | SEQ ID NO 229 | NM_015641 | SEQ ID NO 1574 |
| | AK001492 | SEQ ID NO 230 | NM_015678 | SEQ ID NO 1575 |
| 25 | AK001499 | SEQ ID NO 231 | NM_015721 | SEQ ID NO 1576 |
| | AK001630 | SEQ ID NO 232 | NM_015892 | SEQ ID NO 1578 |
| | AK001872 | SEQ ID NO 234 | NM_015895 | SEQ ID NO 1579 |
| | AK001890 | SEQ ID NO 235 | NM_015907 | SEQ ID NO 1580 |
| | AK002016 | SEQ ID NO 236 | NM_015925 | SEQ ID NO 1581 |
| | AK002088 | SEQ ID NO 237 | NM_015937 | SEQ ID NO 1582 |
| 30 | AK002206 | SEQ ID NO 240 | NM_015954 | SEQ ID NO 1583 |
| | AL035297 | SEQ ID NO 241 | NM_015955 | SEQ ID NO 1584 |
| | AL049265 | SEQ ID NO 242 | NM_015961 | SEQ ID NO 1585 |
| | AL049365 | SEQ ID NO 244 | NM_015984 | SEQ ID NO 1587 |
| | AL049370 | SEQ ID NO 245 | NM_015986 | SEQ ID NO 1588 |
| 35 | AL049381 | SEQ ID NO 246 | NM_015987 | SEQ ID NO 1589 |
| | AL049397 | SEQ ID NO 247 | NM_015991 | SEQ ID NO 1590 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|---------------|-----------------------------|----------------|
| | AL049415 | SEQ ID NO 248 | NM_016002 | SEQ ID NO 1592 |
| | AL049667 | SEQ ID NO 249 | NM_016028 | SEQ ID NO 1594 |
| 5 | AL049801 | SEQ ID NO 250 | NM_016029 | SEQ ID NO 1595 |
| | AL049932 | SEQ ID NO 251 | NM_016047 | SEQ ID NO 1596 |
| | AL049935 | SEQ ID NO 252 | NM_016048 | SEQ ID NO 1597 |
| | AL049943 | SEQ ID NO 253 | NM_016050 | SEQ ID NO 1598 |
| | AL049949 | SEQ ID NO 254 | NM_016056 | SEQ ID NO 1599 |
| | AL049963 | SEQ ID NO 255 | NM_016058 | SEQ ID NO 1600 |
| 10 | AL049987 | SEQ ID NO 256 | NM_016066 | SEQ ID NO 1601 |
| | AL050021 | SEQ ID NO 257 | NM_016072 | SEQ ID NO 1602 |
| | AL050024 | SEQ ID NO 258 | NM_016073 | SEQ ID NO 1603 |
| | AL050090 | SEQ ID NO 259 | NM_016108 | SEQ ID NO 1605 |
| | AL050148 | SEQ ID NO 260 | NM_016109 | SEQ ID NO 1606 |
| 15 | AL050151 | SEQ ID NO 261 | NM_016121 | SEQ ID NO 1607 |
| | AL050227 | SEQ ID NO 262 | NM_016126 | SEQ ID NO 1608 |
| | AL050367 | SEQ ID NO 263 | NM_016127 | SEQ ID NO 1609 |
| | AL050370 | SEQ ID NO 264 | NM_016135 | SEQ ID NO 1610 |
| | AL050371 | SEQ ID NO 265 | NM_016142 | SEQ ID NO 1612 |
| | AL050372 | SEQ ID NO 266 | NM_016153 | SEQ ID NO 1613 |
| 20 | AL050388 | SEQ ID NO 267 | NM_016171 | SEQ ID NO 1614 |
| | AL079276 | SEQ ID NO 268 | NM_016175 | SEQ ID NO 1615 |
| | AL079298 | SEQ ID NO 269 | NM_016184 | SEQ ID NO 1616 |
| | AL080079 | SEQ ID NO 271 | NM_016185 | SEQ ID NO 1617 |
| | AL080192 | SEQ ID NO 273 | NM_016187 | SEQ ID NO 1618 |
| 25 | AL080199 | SEQ ID NO 274 | NM_016199 | SEQ ID NO 1619 |
| | AL080209 | SEQ ID NO 275 | NM_016210 | SEQ ID NO 1620 |
| | AL080234 | SEQ ID NO 277 | NM_016217 | SEQ ID NO 1621 |
| | AL080235 | SEQ ID NO 278 | NM_016228 | SEQ ID NO 1623 |
| | AL096737 | SEQ ID NO 279 | NM_016229 | SEQ ID NO 1624 |
| | AL110126 | SEQ ID NO 280 | NM_016235 | SEQ ID NO 1625 |
| 30 | AL110139 | SEQ ID NO 281 | NM_016240 | SEQ ID NO 1626 |
| | AL110202 | SEQ ID NO 283 | NM_016243 | SEQ ID NO 1627 |
| | AL110212 | SEQ ID NO 284 | NM_016250 | SEQ ID NO 1628 |
| | AL110260 | SEQ ID NO 285 | NM_016267 | SEQ ID NO 1629 |
| | AL117441 | SEQ ID NO 286 | NM_016271 | SEQ ID NO 1630 |
| 35 | AL117452 | SEQ ID NO 287 | NM_016299 | SEQ ID NO 1631 |
| | AL117477 | SEQ ID NO 288 | NM_016306 | SEQ ID NO 1632 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|---------------|-----------------------------|----------------|
| | AL117502 | SEQ ID NO 289 | NM_016308 | SEQ ID NO 1634 |
| | AL117523 | SEQ ID NO 290 | NM_016321 | SEQ ID NO 1635 |
| 5 | AL117595 | SEQ ID NO 291 | NM_016337 | SEQ ID NO 1636 |
| | AL117599 | SEQ ID NO 292 | NM_016352 | SEQ ID NO 1637 |
| | AL117600 | SEQ ID NO 293 | NM_016359 | SEQ ID NO 1638 |
| | AL117609 | SEQ ID NO 294 | NM_016401 | SEQ ID NO 1641 |
| | AL117617 | SEQ ID NO 295 | NM_016403 | SEQ ID NO 1642 |
| | AL117666 | SEQ ID NO 296 | NM_016411 | SEQ ID NO 1643 |
| 10 | AL122055 | SEQ ID NO 297 | NM_016423 | SEQ ID NO 1644 |
| | AL133033 | SEQ ID NO 298 | NM_016463 | SEQ ID NO 1647 |
| | AL133035 | SEQ ID NO 299 | NM_016475 | SEQ ID NO 1649 |
| | AL133074 | SEQ ID NO 301 | NM_016477 | SEQ ID NO 1650 |
| | AL133096 | SEQ ID NO 302 | NM_016491 | SEQ ID NO 1651 |
| 15 | AL133105 | SEQ ID NO 303 | NM_016495 | SEQ ID NO 1652 |
| | AL133108 | SEQ ID NO 304 | NM_016542 | SEQ ID NO 1653 |
| | AL133572 | SEQ ID NO 305 | NM_016548 | SEQ ID NO 1654 |
| | AL133619 | SEQ ID NO 307 | NM_016569 | SEQ ID NO 1655 |
| | AL133622 | SEQ ID NO 308 | NM_016577 | SEQ ID NO 1656 |
| | AL133623 | SEQ ID NO 309 | NM_016582 | SEQ ID NO 1657 |
| 20 | AL133624 | SEQ ID NO 310 | NM_016593 | SEQ ID NO 1658 |
| | AL133632 | SEQ ID NO 311 | NM_016603 | SEQ ID NO 1659 |
| | AL133644 | SEQ ID NO 312 | NM_016612 | SEQ ID NO 1660 |
| | AL133645 | SEQ ID NO 313 | NM_016619 | SEQ ID NO 1661 |
| | AL133651 | SEQ ID NO 314 | NM_016623 | SEQ ID NO 1663 |
| 25 | AL137310 | SEQ ID NO 316 | NM_016625 | SEQ ID NO 1664 |
| | AL137316 | SEQ ID NO 317 | NM_016629 | SEQ ID NO 1665 |
| | AL137332 | SEQ ID NO 318 | NM_016640 | SEQ ID NO 1666 |
| | AL137342 | SEQ ID NO 319 | NM_016645 | SEQ ID NO 1667 |
| | AL137362 | SEQ ID NO 321 | NM_016650 | SEQ ID NO 1668 |
| | AL137381 | SEQ ID NO 322 | NM_016657 | SEQ ID NO 1669 |
| 30 | AL137407 | SEQ ID NO 323 | NM_016733 | SEQ ID NO 1670 |
| | AL137448 | SEQ ID NO 324 | NM_016815 | SEQ ID NO 1671 |
| | AL137502 | SEQ ID NO 326 | NM_016817 | SEQ ID NO 1672 |
| | AL137514 | SEQ ID NO 327 | NM_016818 | SEQ ID NO 1673 |
| | AL137540 | SEQ ID NO 328 | NM_016839 | SEQ ID NO 1675 |
| 35 | AL137566 | SEQ ID NO 330 | NM_017414 | SEQ ID NO 1676 |
| | AL137615 | SEQ ID NO 331 | NM_017422 | SEQ ID NO 1677 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|---------------|-----------------------------|----------------|
| | AL137673 | SEQ ID NO 335 | NM_017423 | SEQ ID NO 1678 |
| | AL137718 | SEQ ID NO 336 | NM_017447 | SEQ ID NO 1679 |
| 5 | AL137736 | SEQ ID NO 337 | NM_017518 | SEQ ID NO 1680 |
| | AL137751 | SEQ ID NO 338 | NM_017522 | SEQ ID NO 1681 |
| | AL137761 | SEQ ID NO 339 | NM_017540 | SEQ ID NO 1682 |
| | AL157431 | SEQ ID NO 340 | NM_017555 | SEQ ID NO 1683 |
| | AL157432 | SEQ ID NO 341 | NM_017572 | SEQ ID NO 1684 |
| | AL157454 | SEQ ID NO 342 | NM_017585 | SEQ ID NO 1685 |
| 10 | AL157476 | SEQ ID NO 343 | NM_017586 | SEQ ID NO 1686 |
| | AL157480 | SEQ ID NO 344 | NM_017596 | SEQ ID NO 1687 |
| | AL157482 | SEQ ID NO 345 | NM_017606 | SEQ ID NO 1688 |
| | AL157484 | SEQ ID NO 346 | NM_017617 | SEQ ID NO 1689 |
| | AL157492 | SEQ ID NO 347 | NM_017633 | SEQ ID NO 1690 |
| 15 | AL157505 | SEQ ID NO 348 | NM_017634 | SEQ ID NO 1691 |
| | AL157851 | SEQ ID NO 349 | NM_017646 | SEQ ID NO 1692 |
| | AL160131 | SEQ ID NO 350 | NM_017660 | SEQ ID NO 1693 |
| | AL161960 | SEQ ID NO 351 | NM_017680 | SEQ ID NO 1694 |
| | AL162049 | SEQ ID NO 352 | NM_017691 | SEQ ID NO 1695 |
| | AL355708 | SEQ ID NO 353 | NM_017698 | SEQ ID NO 1696 |
| 20 | D13643 | SEQ ID NO 355 | NM_017702 | SEQ ID NO 1697 |
| | D14678 | SEQ ID NO 356 | NM_017731 | SEQ ID NO 1699 |
| | D25328 | SEQ ID NO 357 | NM_017732 | SEQ ID NO 1700 |
| | D26070 | SEQ ID NO 358 | NM_017733 | SEQ ID NO 1701 |
| | D26488 | SEQ ID NO 359 | NM_017734 | SEQ ID NO 1702 |
| 25 | D31887 | SEQ ID NO 360 | NM_017746 | SEQ ID NO 1703 |
| | D38521 | SEQ ID NO 361 | NM_017750 | SEQ ID NO 1704 |
| | D38553 | SEQ ID NO 362 | NM_017761 | SEQ ID NO 1705 |
| | D42043 | SEQ ID NO 363 | NM_017763 | SEQ ID NO 1706 |
| | D42047 | SEQ ID NO 364 | NM_017770 | SEQ ID NO 1707 |
| | D43950 | SEQ ID NO 365 | NM_017779 | SEQ ID NO 1708 |
| 30 | D50402 | SEQ ID NO 366 | NM_017780 | SEQ ID NO 1709 |
| | D50914 | SEQ ID NO 367 | NM_017782 | SEQ ID NO 1710 |
| | D55716 | SEQ ID NO 368 | NM_017786 | SEQ ID NO 1711 |
| | D80001 | SEQ ID NO 369 | NM_017791 | SEQ ID NO 1712 |
| | D80010 | SEQ ID NO 370 | NM_017805 | SEQ ID NO 1713 |
| 35 | D82345 | SEQ ID NO 371 | NM_017816 | SEQ ID NO 1714 |
| | D83781 | SEQ ID NO 372 | NM_017821 | SEQ ID NO 1715 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|---------------|-----------------------------|----------------|
| | D86964 | SEQ ID NO 373 | NM_017835 | SEQ ID NO 1716 |
| | D86978 | SEQ ID NO 374 | NM_017843 | SEQ ID NO 1717 |
| 5 | D86985 | SEQ ID NO 375 | NM_017857 | SEQ ID NO 1718 |
| | D87076 | SEQ ID NO 376 | NM_017901 | SEQ ID NO 1719 |
| | D87453 | SEQ ID NO 377 | NM_017906 | SEQ ID NO 1720 |
| | D87469 | SEQ ID NO 378 | NM_017918 | SEQ ID NO 1721 |
| | D87682 | SEQ ID NO 379 | NM_017961 | SEQ ID NO 1722 |
| | G26403 | SEQ ID NO 380 | NM_017996 | SEQ ID NO 1723 |
| 10 | J02639 | SEQ ID NO 381 | NM_018000 | SEQ ID NO 1724 |
| | J04162 | SEQ ID NO 382 | NM_018004 | SEQ ID NO 1725 |
| | K02403 | SEQ ID NO 384 | NM_018011 | SEQ ID NO 1726 |
| | L05096 | SEQ ID NO 385 | NM_018014 | SEQ ID NO 1727 |
| | L10333 | SEQ ID NO 386 | NM_018022 | SEQ ID NO 1728 |
| 15 | L11645 | SEQ ID NO 387 | NM_018031 | SEQ ID NO 1729 |
| | L21934 | SEQ ID NO 388 | NM_018043 | SEQ ID NO 1730 |
| | L22005 | SEQ ID NO 389 | NM_018048 | SEQ ID NO 1731 |
| | L48692 | SEQ ID NO 391 | NM_018062 | SEQ ID NO 1732 |
| | M12758 | SEQ ID NO 392 | NM_018069 | SEQ ID NO 1733 |
| | M15178 | SEQ ID NO 393 | NM_018072 | SEQ ID NO 1734 |
| 20 | M21551 | SEQ ID NO 394 | NM_018077 | SEQ ID NO 1735 |
| | M24895 | SEQ ID NO 395 | NM_018086 | SEQ ID NO 1736 |
| | M26383 | SEQ ID NO 396 | NM_018087 | SEQ ID NO 1737 |
| | M27749 | SEQ ID NO 397 | NM_018093 | SEQ ID NO 1738 |
| | M28170 | SEQ ID NO 398 | NM_018098 | SEQ ID NO 1739 |
| 25 | M29873 | SEQ ID NO 399 | NM_018099 | SEQ ID NO 1740 |
| | M29874 | SEQ ID NO 400 | NM_018101 | SEQ ID NO 1741 |
| | M30448 | SEQ ID NO 401 | NM_018103 | SEQ ID NO 1742 |
| | M30818 | SEQ ID NO 402 | NM_018109 | SEQ ID NO 1744 |
| | M31932 | SEQ ID NO 403 | NM_018123 | SEQ ID NO 1746 |
| | M37033 | SEQ ID NO 404 | NM_018131 | SEQ ID NO 1747 |
| 30 | M55914 | SEQ ID NO 405 | NM_018136 | SEQ ID NO 1748 |
| | M63438 | SEQ ID NO 406 | NM_018138 | SEQ ID NO 1749 |
| | M65254 | SEQ ID NO 407 | NM_018166 | SEQ ID NO 1750 |
| | M68874 | SEQ ID NO 408 | NM_018171 | SEQ ID NO 1751 |
| | M73547 | SEQ ID NO 409 | NM_018178 | SEQ ID NO 1752 |
| 35 | M77142 | SEQ ID NO 410 | NM_018181 | SEQ ID NO 1753 |
| | M80899 | SEQ ID NO 411 | NM_018186 | SEQ ID NO 1754 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|---------------|-----------------------------|----------------|
| | M83822 | SEQ ID NO 412 | NM_018188 | SEQ ID NO 1756 |
| | M90657 | SEQ ID NO 413 | NM_018194 | SEQ ID NO 1757 |
| 5 | M93718 | SEQ ID NO 414 | NM_018204 | SEQ ID NO 1758 |
| | M96577 | SEQ ID NO 415 | NM_018208 | SEQ ID NO 1759 |
| | NM_000022 | SEQ ID NO 417 | NM_018212 | SEQ ID NO 1760 |
| | NM_000044 | SEQ ID NO 418 | NM_018234 | SEQ ID NO 1763 |
| | NM_000050 | SEQ ID NO 419 | NM_018255 | SEQ ID NO 1764 |
| | NM_000057 | SEQ ID NO 420 | NM_018257 | SEQ ID NO 1765 |
| 10 | NM_000060 | SEQ ID NO 421 | NM_018265 | SEQ ID NO 1766 |
| | NM_000064 | SEQ ID NO 422 | NM_018271 | SEQ ID NO 1767 |
| | NM_000073 | SEQ ID NO 424 | NM_018290 | SEQ ID NO 1768 |
| | NM_000077 | SEQ ID NO 425 | NM_018295 | SEQ ID NO 1769 |
| | NM_000086 | SEQ ID NO 426 | NM_018304 | SEQ ID NO 1770 |
| 15 | NM_000087 | SEQ ID NO 427 | NM_018306 | SEQ ID NO 1771 |
| | NM_000095 | SEQ ID NO 429 | NM_018326 | SEQ ID NO 1772 |
| | NM_000096 | SEQ ID NO 430 | NM_018346 | SEQ ID NO 1773 |
| | NM_000100 | SEQ ID NO 431 | NM_018366 | SEQ ID NO 1775 |
| | NM_000101 | SEQ ID NO 432 | NM_018370 | SEQ ID NO 1776 |
| | NM_000104 | SEQ ID NO 433 | NM_018373 | SEQ ID NO 1777 |
| 20 | NM_000109 | SEQ ID NO 434 | NM_018379 | SEQ ID NO 1778 |
| | NM_000125 | SEQ ID NO 435 | NM_018384 | SEQ ID NO 1779 |
| | NM_000127 | SEQ ID NO 436 | NM_018389 | SEQ ID NO 1780 |
| | NM_000135 | SEQ ID NO 437 | NM_018410 | SEQ ID NO 1783 |
| | NM_000137 | SEQ ID NO 438 | NM_018439 | SEQ ID NO 1785 |
| 25 | NM_000146 | SEQ ID NO 439 | NM_018454 | SEQ ID NO 1786 |
| | NM_000149 | SEQ ID NO 440 | NM_018455 | SEQ ID NO 1787 |
| | NM_000154 | SEQ ID NO 441 | NM_018465 | SEQ ID NO 1788 |
| | NM_000161 | SEQ ID NO 443 | NM_018471 | SEQ ID NO 1789 |
| | NM_000165 | SEQ ID NO 444 | NM_018478 | SEQ ID NO 1790 |
| | NM_000168 | SEQ ID NO 445 | NM_018479 | SEQ ID NO 1791 |
| 30 | NM_000169 | SEQ ID NO 446 | NM_018529 | SEQ ID NO 1793 |
| | NM_000175 | SEQ ID NO 447 | NM_018556 | SEQ ID NO 1794 |
| | NM_000191 | SEQ ID NO 448 | NM_018569 | SEQ ID NO 1795 |
| | NM_000201 | SEQ ID NO 450 | NM_018584 | SEQ ID NO 1796 |
| | NM_000211 | SEQ ID NO 451 | NM_018653 | SEQ ID NO 1797 |
| 35 | NM_000213 | SEQ ID NO 452 | NM_018660 | SEQ ID NO 1798 |
| | NM_000224 | SEQ ID NO 453 | NM_018683 | SEQ ID NO 1799 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|---------------|-----------------------------|----------------|
| | NM_000239 | SEQ ID NO 454 | NM_018685 | SEQ ID NO 1800 |
| | NM_000251 | SEQ ID NO 455 | NM_018686 | SEQ ID NO 1801 |
| 5 | NM_000268 | SEQ ID NO 456 | NM_018695 | SEQ ID NO 1802 |
| | NM_000270 | SEQ ID NO 458 | NM_018728 | SEQ ID NO 1803 |
| | NM_000271 | SEQ ID NO 459 | NM_018840 | SEQ ID NO 1804 |
| | NM_000283 | SEQ ID NO 460 | NM_018842 | SEQ ID NO 1805 |
| | NM_000284 | SEQ ID NO 461 | NM_018950 | SEQ ID NO 1806 |
| | NM_000286 | SEQ ID NO 462 | NM_018988 | SEQ ID NO 1807 |
| 10 | NM_000291 | SEQ ID NO 463 | NM_019000 | SEQ ID NO 1808 |
| | NM_000299 | SEQ ID NO 464 | NM_019013 | SEQ ID NO 1809 |
| | NM_000300 | SEQ ID NO 465 | NM_019025 | SEQ ID NO 1810 |
| | NM_000310 | SEQ ID NO 466 | NM_019027 | SEQ ID NO 1811 |
| | NM_000311 | SEQ ID NO 467 | NM_019041 | SEQ ID NO 1812 |
| 15 | NM_000317 | SEQ ID NO 468 | NM_019044 | SEQ ID NO 1813 |
| | NM_000320 | SEQ ID NO 469 | NM_019063 | SEQ ID NO 1815 |
| | NM_000342 | SEQ ID NO 470 | NM_019084 | SEQ ID NO 1816 |
| | NM_000346 | SEQ ID NO 471 | NM_019554 | SEQ ID NO 1817 |
| | NM_000352 | SEQ ID NO 472 | NM_019845 | SEQ ID NO 1818 |
| | NM_000355 | SEQ ID NO 473 | NM_019858 | SEQ ID NO 1819 |
| 20 | NM_000358 | SEQ ID NO 474 | NM_020130 | SEQ ID NO 1820 |
| | NM_000359 | SEQ ID NO 475 | NM_020133 | SEQ ID NO 1821 |
| | NM_000362 | SEQ ID NO 476 | NM_020143 | SEQ ID NO 1822 |
| | NM_000365 | SEQ ID NO 477 | NM_020150 | SEQ ID NO 1823 |
| | NM_000381 | SEQ ID NO 478 | NM_020163 | SEQ ID NO 1824 |
| 25 | NM_000397 | SEQ ID NO 480 | NM_020166 | SEQ ID NO 1825 |
| | NM_000399 | SEQ ID NO 481 | NM_020169 | SEQ ID NO 1826 |
| | NM_000414 | SEQ ID NO 482 | NM_020179 | SEQ ID NO 1827 |
| | NM_000416 | SEQ ID NO 483 | NM_020184 | SEQ ID NO 1828 |
| | NM_000422 | SEQ ID NO 484 | NM_020186 | SEQ ID NO 1829 |
| | NM_000424 | SEQ ID NO 485 | NM_020188 | SEQ ID NO 1830 |
| 30 | NM_000433 | SEQ ID NO 486 | NM_020189 | SEQ ID NO 1831 |
| | NM_000436 | SEQ ID NO 487 | NM_020197 | SEQ ID NO 1832 |
| | NM_000450 | SEQ ID NO 488 | NM_020199 | SEQ ID NO 1833 |
| | NM_000462 | SEQ ID NO 489 | NM_020215 | SEQ ID NO 1834 |
| | NM_000495 | SEQ ID NO 490 | NM_020347 | SEQ ID NO 1836 |
| 35 | NM_000507 | SEQ ID NO 491 | NM_020365 | SEQ ID NO 1837 |
| | NM_000526 | SEQ ID NO 492 | NM_020386 | SEQ ID NO 1838 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|---------------|-----------------------------|----------------|
| | NM_000557 | SEQ ID NO 493 | NM_020445 | SEQ ID NO 1839 |
| | NM_000560 | SEQ ID NO 494 | NM_020639 | SEQ ID NO 1840 |
| 5 | NM_000576 | SEQ ID NO 495 | NM_020659 | SEQ ID NO 1841 |
| | NM_000579 | SEQ ID NO 496 | NM_020675 | SEQ ID NO 1842 |
| | NM_000584 | SEQ ID NO 497 | NM_020686 | SEQ ID NO 1843 |
| | NM_000591 | SEQ ID NO 498 | NM_020974 | SEQ ID NO 1844 |
| | NM_000592 | SEQ ID NO 499 | NM_020978 | SEQ ID NO 1845 |
| | NM_000593 | SEQ ID NO 500 | NM_020979 | SEQ ID NO 1846 |
| 10 | NM_000594 | SEQ ID NO 501 | NM_020980 | SEQ ID NO 1847 |
| | NM_000597 | SEQ ID NO 502 | NM_021000 | SEQ ID NO 1849 |
| | NM_000600 | SEQ ID NO 504 | NM_021004 | SEQ ID NO 1850 |
| | NM_000607 | SEQ ID NO 505 | NM_021025 | SEQ ID NO 1851 |
| | NM_000612 | SEQ ID NO 506 | NM_021063 | SEQ ID NO 1852 |
| 15 | NM_000627 | SEQ ID NO 507 | NM_021065 | SEQ ID NO 1853 |
| | NM_000633 | SEQ ID NO 508 | NM_021077 | SEQ ID NO 1854 |
| | NM_000636 | SEQ ID NO 509 | NM_021095 | SEQ ID NO 1855 |
| | NM_000639 | SEQ ID NO 510 | NM_021101 | SEQ ID NO 1856 |
| | NM_000647 | SEQ ID NO 511 | NM_021103 | SEQ ID NO 1857 |
| | NM_000655 | SEQ ID NO 512 | NM_021128 | SEQ ID NO 1858 |
| 20 | NM_000662 | SEQ ID NO 513 | NM_021147 | SEQ ID NO 1859 |
| | NM_000663 | SEQ ID NO 514 | NM_021151 | SEQ ID NO 1860 |
| | NM_000666 | SEQ ID NO 515 | NM_021181 | SEQ ID NO 1861 |
| | NM_000676 | SEQ ID NO 516 | NM_021190 | SEQ ID NO 1862 |
| | NM_000685 | SEQ ID NO 517 | NM_021198 | SEQ ID NO 1863 |
| 25 | NM_000693 | SEQ ID NO 518 | NM_021200 | SEQ ID NO 1864 |
| | NM_000699 | SEQ ID NO 519 | NM_021203 | SEQ ID NO 1865 |
| | NM_000700 | SEQ ID NO 520 | NM_021238 | SEQ ID NO 1866 |
| | NM_000712 | SEQ ID NO 521 | NM_021242 | SEQ ID NO 1867 |
| | NM_000727 | SEQ ID NO 522 | S40706 | SEQ ID NO 1869 |
| | NM_000732 | SEQ ID NO 523 | S53354 | SEQ ID NO 1870 |
| 30 | NM_000734 | SEQ ID NO 524 | S59184 | SEQ ID NO 1871 |
| | NM_000767 | SEQ ID NO 525 | S62138 | SEQ ID NO 1872 |
| | NM_000784 | SEQ ID NO 526 | U09848 | SEQ ID NO 1873 |
| | NM_000802 | SEQ ID NO 528 | U10991 | SEQ ID NO 1874 |
| | NM_000824 | SEQ ID NO 529 | U17077 | SEQ ID NO 1875 |
| 35 | NM_000849 | SEQ ID NO 530 | U18919 | SEQ ID NO 1876 |
| | NM_000852 | SEQ ID NO 531 | U41387 | SEQ ID NO 1877 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|---------------|-----------------------------|----------------|
| | NM_000874 | SEQ ID NO 532 | U45975 | SEQ ID NO 1878 |
| | NM_000878 | SEQ ID NO 533 | U49835 | SEQ ID NO 1879 |
| 5 | NM_000884 | SEQ ID NO 534 | U56725 | SEQ ID NO 1880 |
| | NM_000908 | SEQ ID NO 537 | U58033 | SEQ ID NO 1881 |
| | NM_000909 | SEQ ID NO 538 | U61167 | SEQ ID NO 1882 |
| | NM_000926 | SEQ ID NO 539 | U66042 | SEQ ID NO 1883 |
| | NM_000930 | SEQ ID NO 540 | U68385 | SEQ ID NO 1885 |
| | NM_000931 | SEQ ID NO 541 | U68494 | SEQ ID NO 1886 |
| 10 | NM_000947 | SEQ ID NO 542 | U74612 | SEQ ID NO 1887 |
| | NM_000949 | SEQ ID NO 543 | U75968 | SEQ ID NO 1888 |
| | NM_000950 | SEQ ID NO 544 | U79293 | SEQ ID NO 1889 |
| | NM_000954 | SEQ ID NO 545 | U80736 | SEQ ID NO 1890 |
| | NM_000964 | SEQ ID NO 546 | U82987 | SEQ ID NO 1891 |
| 15 | NM_001003 | SEQ ID NO 549 | U83115 | SEQ ID NO 1892 |
| | NM_001016 | SEQ ID NO 551 | U89715 | SEQ ID NO 1893 |
| | NM_001047 | SEQ ID NO 553 | U90916 | SEQ ID NO 1894 |
| | NM_001066 | SEQ ID NO 555 | U92544 | SEQ ID NO 1895 |
| | NM_001071 | SEQ ID NO 556 | U96131 | SEQ ID NO 1896 |
| | NM_001078 | SEQ ID NO 557 | U96394 | SEQ ID NO 1897 |
| 20 | NM_001085 | SEQ ID NO 558 | W61000_RC | SEQ ID NO 1898 |
| | NM_001089 | SEQ ID NO 559 | X00437 | SEQ ID NO 1899 |
| | NM_001109 | SEQ ID NO 560 | X00497 | SEQ ID NO 1900 |
| | NM_001122 | SEQ ID NO 561 | X01394 | SEQ ID NO 1901 |
| | NM_001124 | SEQ ID NO 562 | X03084 | SEQ ID NO 1902 |
| 25 | NM_001161 | SEQ ID NO 563 | X07834 | SEQ ID NO 1905 |
| | NM_001165 | SEQ ID NO 564 | X14356 | SEQ ID NO 1906 |
| | NM_001166 | SEQ ID NO 565 | X16302 | SEQ ID NO 1907 |
| | NM_001168 | SEQ ID NO 566 | X52486 | SEQ ID NO 1909 |
| | NM_001179 | SEQ ID NO 567 | X52882 | SEQ ID NO 1910 |
| | NM_001185 | SEQ ID NO 569 | X56807 | SEQ ID NO 1911 |
| 30 | NM_001203 | SEQ ID NO 570 | X57809 | SEQ ID NO 1912 |
| | NM_001207 | SEQ ID NO 573 | X57819 | SEQ ID NO 1913 |
| | NM_001216 | SEQ ID NO 574 | X58529 | SEQ ID NO 1914 |
| | NM_001218 | SEQ ID NO 575 | X59405 | SEQ ID NO 1915 |
| | NM_001223 | SEQ ID NO 576 | X72475 | SEQ ID NO 1918 |
| 35 | NM_001225 | SEQ ID NO 577 | X73617 | SEQ ID NO 1919 |
| | NM_001233 | SEQ ID NO 578 | X74794 | SEQ ID NO 1920 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|---------------|-----------------------------|----------------|
| | NM_001236 | SEQ ID NO 579 | X75315 | SEQ ID NO 1921 |
| | NM_001237 | SEQ ID NO 580 | X79782 | SEQ ID NO 1922 |
| 5 | NM_001251 | SEQ ID NO 581 | X82693 | SEQ ID NO 1923 |
| | NM_001255 | SEQ ID NO 582 | X83301 | SEQ ID NO 1924 |
| | NM_001262 | SEQ ID NO 583 | X93006 | SEQ ID NO 1926 |
| | NM_001263 | SEQ ID NO 584 | X94232 | SEQ ID NO 1927 |
| | NM_001267 | SEQ ID NO 585 | X98834 | SEQ ID NO 1929 |
| | NM_001276 | SEQ ID NO 587 | X99142 | SEQ ID NO 1930 |
| 10 | NM_001280 | SEQ ID NO 588 | Y14737 | SEQ ID NO 1932 |
| | NM_001282 | SEQ ID NO 589 | Z11887 | SEQ ID NO 1933 |
| | NM_001295 | SEQ ID NO 590 | Z48633 | SEQ ID NO 1935 |
| | NM_001305 | SEQ ID NO 591 | NM_004222 | SEQ ID NO 1936 |
| | NM_001310 | SEQ ID NO 592 | NM_016405 | SEQ ID NO 1937 |
| 15 | NM_001312 | SEQ ID NO 593 | NM_017690 | SEQ ID NO 1938 |
| | NM_001321 | SEQ ID NO 594 | Contig29_RC | SEQ ID NO 1939 |
| | NM_001327 | SEQ ID NO 595 | Contig237_RC | SEQ ID NO 1940 |
| | NM_001329 | SEQ ID NO 596 | Contig263_RC | SEQ ID NO 1941 |
| | NM_001333 | SEQ ID NO 597 | Contig292_RC | SEQ ID NO 1942 |
| | NM_001338 | SEQ ID NO 598 | Contig382_RC | SEQ ID NO 1944 |
| 20 | NM_001360 | SEQ ID NO 599 | Contig399_RC | SEQ ID NO 1945 |
| | NM_001363 | SEQ ID NO 600 | Contig448_RC | SEQ ID NO 1946 |
| | NM_001381 | SEQ ID NO 601 | Contig569_RC | SEQ ID NO 1947 |
| | NM_001394 | SEQ ID NO 602 | Contig580_RC | SEQ ID NO 1948 |
| | NM_001395 | SEQ ID NO 603 | Contig678_RC | SEQ ID NO 1949 |
| 25 | NM_001419 | SEQ ID NO 604 | Contig706_RC | SEQ ID NO 1950 |
| | NM_001424 | SEQ ID NO 605 | Contig718_RC | SEQ ID NO 1951 |
| | NM_001428 | SEQ ID NO 606 | Contig719_RC | SEQ ID NO 1952 |
| | NM_001436 | SEQ ID NO 607 | Contig742_RC | SEQ ID NO 1953 |
| | NM_001444 | SEQ ID NO 608 | Contig753_RC | SEQ ID NO 1954 |
| | NM_001446 | SEQ ID NO 609 | Contig758_RC | SEQ ID NO 1956 |
| 30 | NM_001453 | SEQ ID NO 611 | Contig760_RC | SEQ ID NO 1957 |
| | NM_001456 | SEQ ID NO 612 | Contig842_RC | SEQ ID NO 1958 |
| | NM_001457 | SEQ ID NO 613 | Contig848_RC | SEQ ID NO 1959 |
| | NM_001463 | SEQ ID NO 614 | Contig924_RC | SEQ ID NO 1960 |
| | NM_001465 | SEQ ID NO 615 | Contig974_RC | SEQ ID NO 1961 |
| 35 | NM_001481 | SEQ ID NO 616 | Contig1018_RC | SEQ ID NO 1962 |
| | NM_001493 | SEQ ID NO 617 | Contig1056_RC | SEQ ID NO 1963 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|---------------|-----------------------------|----------------|
| | NM_001494 | SEQ ID NO 618 | Contig1061_RC | SEQ ID NO 1964 |
| | NM_001500 | SEQ ID NO 619 | Contig1129_RC | SEQ ID NO 1965 |
| 5 | NM_001504 | SEQ ID NO 620 | Contig1148 | SEQ ID NO 1966 |
| | NM_001511 | SEQ ID NO 621 | Contig1239_RC | SEQ ID NO 1967 |
| | NM_001513 | SEQ ID NO 622 | Contig1277 | SEQ ID NO 1968 |
| | NM_001527 | SEQ ID NO 623 | Contig1333_RC | SEQ ID NO 1969 |
| | NM_001529 | SEQ ID NO 624 | Contig1386_RC | SEQ ID NO 1970 |
| | NM_001530 | SEQ ID NO 625 | Contig1389_RC | SEQ ID NO 1971 |
| 10 | NM_001540 | SEQ ID NO 626 | Contig1418_RC | SEQ ID NO 1972 |
| | NM_001550 | SEQ ID NO 627 | Contig1462_RC | SEQ ID NO 1973 |
| | NM_001551 | SEQ ID NO 628 | Contig1505_RC | SEQ ID NO 1974 |
| | NM_001552 | SEQ ID NO 629 | Contig1540_RC | SEQ ID NO 1975 |
| | NM_001554 | SEQ ID NO 631 | Contig1584_RC | SEQ ID NO 1976 |
| 15 | NM_001558 | SEQ ID NO 632 | Contig1632_RC | SEQ ID NO 1977 |
| | NM_001560 | SEQ ID NO 633 | Contig1682_RC | SEQ ID NO 1978 |
| | NM_001565 | SEQ ID NO 634 | Contig1778_RC | SEQ ID NO 1979 |
| | NM_001569 | SEQ ID NO 635 | Contig1829 | SEQ ID NO 1981 |
| | NM_001605 | SEQ ID NO 636 | Contig1838_RC | SEQ ID NO 1982 |
| | NM_001609 | SEQ ID NO 637 | Contig1938_RC | SEQ ID NO 1983 |
| 20 | NM_001615 | SEQ ID NO 638 | Contig1970_RC | SEQ ID NO 1984 |
| | NM_001623 | SEQ ID NO 639 | Contig1998_RC | SEQ ID NO 1985 |
| | NM_001627 | SEQ ID NO 640 | Contig2099_RC | SEQ ID NO 1986 |
| | NM_001628 | SEQ ID NO 641 | Contig2143_RC | SEQ ID NO 1987 |
| | NM_001630 | SEQ ID NO 642 | Contig2237_RC | SEQ ID NO 1988 |
| 25 | NM_001634 | SEQ ID NO 643 | Contig2429_RC | SEQ ID NO 1990 |
| | NM_001656 | SEQ ID NO 644 | Contig2504_RC | SEQ ID NO 1991 |
| | NM_001673 | SEQ ID NO 645 | Contig2512_RC | SEQ ID NO 1992 |
| | NM_001675 | SEQ ID NO 647 | Contig2575_RC | SEQ ID NO 1993 |
| | NM_001679 | SEQ ID NO 648 | Contig2578_RC | SEQ ID NO 1994 |
| | NM_001689 | SEQ ID NO 649 | Contig2639_RC | SEQ ID NO 1995 |
| 30 | NM_001703 | SEQ ID NO 650 | Contig2647_RC | SEQ ID NO 1996 |
| | NM_001710 | SEQ ID NO 651 | Contig2657_RC | SEQ ID NO 1997 |
| | NM_001725 | SEQ ID NO 652 | Contig2728_RC | SEQ ID NO 1998 |
| | NM_001730 | SEQ ID NO 653 | Contig2745_RC | SEQ ID NO 1999 |
| | NM_001733 | SEQ ID NO 654 | Contig2811_RC | SEQ ID NO 2000 |
| 35 | NM_001734 | SEQ ID NO 655 | Contig2873_RC | SEQ ID NO 2001 |
| | NM_001740 | SEQ ID NO 656 | Contig2883_RC | SEQ ID NO 2002 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|---------------|-----------------------------|----------------|
| | NM_001745 | SEQ ID NO 657 | Contig2915_RC | SEQ ID NO 2003 |
| | NM_001747 | SEQ ID NO 658 | Contig2928_RC | SEQ ID NO 2004 |
| 5 | NM_001756 | SEQ ID NO 659 | Contig3024_RC | SEQ ID NO 2005 |
| | NM_001757 | SEQ ID NO 660 | Contig3094_RC | SEQ ID NO 2006 |
| | NM_001758 | SEQ ID NO 661 | Contig3164_RC | SEQ ID NO 2007 |
| | NM_001762 | SEQ ID NO 662 | Contig3495_RC | SEQ ID NO 2009 |
| | NM_001767 | SEQ ID NO 663 | Contig3607_RC | SEQ ID NO 2010 |
| | NM_001770 | SEQ ID NO 664 | Contig3659_RC | SEQ ID NO 2011 |
| 10 | NM_001777 | SEQ ID NO 665 | Contig3677_RC | SEQ ID NO 2012 |
| | NM_001778 | SEQ ID NO 666 | Contig3682_RC | SEQ ID NO 2013 |
| | NM_001781 | SEQ ID NO 667 | Contig3734_RC | SEQ ID NO 2014 |
| | NM_001786 | SEQ ID NO 668 | Contig3834_RC | SEQ ID NO 2015 |
| | NM_001793 | SEQ ID NO 669 | Contig3876_RC | SEQ ID NO 2016 |
| 15 | NM_001803 | SEQ ID NO 671 | Contig3902_RC | SEQ ID NO 2017 |
| | NM_001806 | SEQ ID NO 672 | Contig3940_RC | SEQ ID NO 2018 |
| | NM_001809 | SEQ ID NO 673 | Contig4380_RC | SEQ ID NO 2019 |
| | NM_001814 | SEQ ID NO 674 | Contig4388_RC | SEQ ID NO 2020 |
| | NM_001826 | SEQ ID NO 675 | Contig4467_RC | SEQ ID NO 2021 |
| | NM_001830 | SEQ ID NO 677 | Contig4949_RC | SEQ ID NO 2023 |
| 20 | NM_001838 | SEQ ID NO 678 | Contig5348_RC | SEQ ID NO 2024 |
| | NM_001839 | SEQ ID NO 679 | Contig5403_RC | SEQ ID NO 2025 |
| | NM_001853 | SEQ ID NO 681 | Contig5716_RC | SEQ ID NO 2026 |
| | NM_001859 | SEQ ID NO 682 | Contig6118_RC | SEQ ID NO 2027 |
| | NM_001861 | SEQ ID NO 683 | Contig6164_RC | SEQ ID NO 2028 |
| 25 | NM_001874 | SEQ ID NO 685 | Contig6181_RC | SEQ ID NO 2029 |
| | NM_001885 | SEQ ID NO 686 | Contig6514_RC | SEQ ID NO 2030 |
| | NM_001892 | SEQ ID NO 688 | Contig6612_RC | SEQ ID NO 2031 |
| | NM_001897 | SEQ ID NO 689 | Contig6881_RC | SEQ ID NO 2032 |
| | NM_001899 | SEQ ID NO 690 | Contig8165_RC | SEQ ID NO 2033 |
| | NM_001905 | SEQ ID NO 691 | Contig8221_RC | SEQ ID NO 2034 |
| 30 | NM_001912 | SEQ ID NO 692 | Contig8347_RC | SEQ ID NO 2035 |
| | NM_001914 | SEQ ID NO 693 | Contig8364_RC | SEQ ID NO 2036 |
| | NM_001919 | SEQ ID NO 694 | Contig8888_RC | SEQ ID NO 2038 |
| | NM_001941 | SEQ ID NO 695 | Contig9259_RC | SEQ ID NO 2039 |
| | NM_001943 | SEQ ID NO 696 | Contig9541_RC | SEQ ID NO 2040 |
| 35 | NM_001944 | SEQ ID NO 697 | Contig10268_RC | SEQ ID NO 2041 |
| | NM_001953 | SEQ ID NO 699 | Contig10363_RC | SEQ ID NO 2042 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|---------------|-----------------------------|----------------|
| | NM_001954 | SEQ ID NO 700 | Contig10437_RC | SEQ ID NO 2043 |
| | NM_001955 | SEQ ID NO 701 | Contig11086_RC | SEQ ID NO 2045 |
| 5 | NM_001956 | SEQ ID NO 702 | Contig11275_RC | SEQ ID NO 2046 |
| | NM_001958 | SEQ ID NO 703 | Contig11648_RC | SEQ ID NO 2047 |
| | NM_001961 | SEQ ID NO 705 | Contig12216_RC | SEQ ID NO 2048 |
| | NM_001970 | SEQ ID NO 706 | Contig12369_RC | SEQ ID NO 2049 |
| | NM_001979 | SEQ ID NO 707 | Contig12814_RC | SEQ ID NO 2050 |
| | NM_001982 | SEQ ID NO 708 | Contig12951_RC | SEQ ID NO 2051 |
| 10 | NM_002017 | SEQ ID NO 710 | Contig13480_RC | SEQ ID NO 2052 |
| | NM_002033 | SEQ ID NO 713 | Contig14284_RC | SEQ ID NO 2053 |
| | NM_002046 | SEQ ID NO 714 | Contig14390_RC | SEQ ID NO 2054 |
| | NM_002047 | SEQ ID NO 715 | Contig14780_RC | SEQ ID NO 2055 |
| | NM_002051 | SEQ ID NO 716 | Contig14954_RC | SEQ ID NO 2056 |
| 15 | NM_002053 | SEQ ID NO 717 | Contig14981_RC | SEQ ID NO 2057 |
| | NM_002061 | SEQ ID NO 718 | Contig15692_RC | SEQ ID NO 2058 |
| | NM_002065 | SEQ ID NO 719 | Contig16192_RC | SEQ ID NO 2059 |
| | NM_002068 | SEQ ID NO 720 | Contig16759_RC | SEQ ID NO 2061 |
| | NM_002077 | SEQ ID NO 722 | Contig16786_RC | SEQ ID NO 2062 |
| | NM_002091 | SEQ ID NO 723 | Contig16905_RC | SEQ ID NO 2063 |
| 20 | NM_002101 | SEQ ID NO 724 | Contig17103_RC | SEQ ID NO 2064 |
| | NM_002106 | SEQ ID NO 725 | Contig17105_RC | SEQ ID NO 2065 |
| | NM_002110 | SEQ ID NO 726 | Contig17248_RC | SEQ ID NO 2066 |
| | NM_002111 | SEQ ID NO 727 | Contig17345_RC | SEQ ID NO 2067 |
| | NM_002115 | SEQ ID NO 728 | Contig18502_RC | SEQ ID NO 2069 |
| 25 | NM_002118 | SEQ ID NO 729 | Contig20156_RC | SEQ ID NO 2071 |
| | NM_002123 | SEQ ID NO 730 | Contig20302_RC | SEQ ID NO 2073 |
| | NM_002131 | SEQ ID NO 731 | Contig20600_RC | SEQ ID NO 2074 |
| | NM_002136 | SEQ ID NO 732 | Contig20617_RC | SEQ ID NO 2075 |
| | NM_002145 | SEQ ID NO 733 | Contig20629_RC | SEQ ID NO 2076 |
| | NM_002164 | SEQ ID NO 734 | Contig20651_RC | SEQ ID NO 2077 |
| 30 | NM_002168 | SEQ ID NO 735 | Contig21130_RC | SEQ ID NO 2078 |
| | NM_002184 | SEQ ID NO 736 | Contig21185_RC | SEQ ID NO 2079 |
| | NM_002185 | SEQ ID NO 737 | Contig21421_RC | SEQ ID NO 2080 |
| | NM_002189 | SEQ ID NO 738 | Contig21787_RC | SEQ ID NO 2081 |
| | NM_002200 | SEQ ID NO 739 | Contig21812_RC | SEQ ID NO 2082 |
| 35 | NM_002201 | SEQ ID NO 740 | Contig22418_RC | SEQ ID NO 2083 |
| | NM_002213 | SEQ ID NO 741 | Contig23085_RC | SEQ ID NO 2084 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|---------------|-----------------------------|----------------|
| | NM_002219 | SEQ ID NO 742 | Contig23454_RC | SEQ ID NO 2085 |
| | NM_002222 | SEQ ID NO 743 | Contig24138_RC | SEQ ID NO 2086 |
| 5 | NM_002239 | SEQ ID NO 744 | Contig24252_RC | SEQ ID NO 2087 |
| | NM_002243 | SEQ ID NO 745 | Contig24655_RC | SEQ ID NO 2089 |
| | NM_002245 | SEQ ID NO 746 | Contig25055_RC | SEQ ID NO 2090 |
| | NM_002250 | SEQ ID NO 747 | Contig25290_RC | SEQ ID NO 2091 |
| | NM_002254 | SEQ ID NO 748 | Contig25343_RC | SEQ ID NO 2092 |
| | NM_002266 | SEQ ID NO 749 | Contig25362_RC | SEQ ID NO 2093 |
| 10 | NM_002273 | SEQ ID NO 750 | Contig25617_RC | SEQ ID NO 2094 |
| | NM_002281 | SEQ ID NO 751 | Contig25659_RC | SEQ ID NO 2095 |
| | NM_002292 | SEQ ID NO 752 | Contig25722_RC | SEQ ID NO 2096 |
| | NM_002298 | SEQ ID NO 753 | Contig25809_RC | SEQ ID NO 2097 |
| | NM_002300 | SEQ ID NO 754 | Contig25991 | SEQ ID NO 2098 |
| 15 | NM_002308 | SEQ ID NO 755 | Contig26022_RC | SEQ ID NO 2099 |
| | NM_002314 | SEQ ID NO 756 | Contig26077_RC | SEQ ID NO 2100 |
| | NM_002337 | SEQ ID NO 757 | Contig26310_RC | SEQ ID NO 2101 |
| | NM_002341 | SEQ ID NO 758 | Contig26371_RC | SEQ ID NO 2102 |
| | NM_002342 | SEQ ID NO 759 | Contig26438_RC | SEQ ID NO 2103 |
| | NM_002346 | SEQ ID NO 760 | Contig26706_RC | SEQ ID NO 2104 |
| 20 | NM_002349 | SEQ ID NO 761 | Contig27088_RC | SEQ ID NO 2105 |
| | NM_002350 | SEQ ID NO 762 | Contig27186_RC | SEQ ID NO 2106 |
| | NM_002356 | SEQ ID NO 763 | Contig27228_RC | SEQ ID NO 2107 |
| | NM_002358 | SEQ ID NO 764 | Contig27344_RC | SEQ ID NO 2109 |
| | NM_002370 | SEQ ID NO 765 | Contig27386_RC | SEQ ID NO 2110 |
| 25 | NM_002395 | SEQ ID NO 766 | Contig27624_RC | SEQ ID NO 2111 |
| | NM_002416 | SEQ ID NO 767 | Contig27749_RC | SEQ ID NO 2112 |
| | NM_002421 | SEQ ID NO 768 | Contig27882_RC | SEQ ID NO 2113 |
| | NM_002426 | SEQ ID NO 769 | Contig27915_RC | SEQ ID NO 2114 |
| | NM_002435 | SEQ ID NO 770 | Contig28030_RC | SEQ ID NO 2115 |
| | NM_002438 | SEQ ID NO 771 | Contig28081_RC | SEQ ID NO 2116 |
| 30 | NM_002444 | SEQ ID NO 772 | Contig28152_RC | SEQ ID NO 2117 |
| | NM_002449 | SEQ ID NO 773 | Contig28550_RC | SEQ ID NO 2119 |
| | NM_002450 | SEQ ID NO 774 | Contig28552_RC | SEQ ID NO 2120 |
| | NM_002456 | SEQ ID NO 775 | Contig28712_RC | SEQ ID NO 2121 |
| | NM_002466 | SEQ ID NO 776 | Contig28888_RC | SEQ ID NO 2122 |
| 35 | NM_002482 | SEQ ID NO 777 | Contig28947_RC | SEQ ID NO 2123 |
| | NM_002497 | SEQ ID NO 778 | Contig29126_RC | SEQ ID NO 2124 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|---------------|-----------------------------|----------------|
| | NM_002510 | SEQ ID NO 779 | Contig29193_RC | SEQ ID NO 2125 |
| | NM_002515 | SEQ ID NO 781 | Contig29369_RC | SEQ ID NO 2126 |
| 5 | NM_002524 | SEQ ID NO 782 | Contig29639_RC | SEQ ID NO 2127 |
| | NM_002539 | SEQ ID NO 783 | Contig30047_RC | SEQ ID NO 2129 |
| | NM_002555 | SEQ ID NO 785 | Contig30154_RC | SEQ ID NO 2131 |
| | NM_002570 | SEQ ID NO 787 | Contig30209_RC | SEQ ID NO 2132 |
| | NM_002579 | SEQ ID NO 788 | Contig30213_RC | SEQ ID NO 2133 |
| | NM_002587 | SEQ ID NO 789 | Contig30230_RC | SEQ ID NO 2134 |
| 10 | NM_002590 | SEQ ID NO 790 | Contig30267_RC | SEQ ID NO 2135 |
| | NM_002600 | SEQ ID NO 791 | Contig30390_RC | SEQ ID NO 2136 |
| | NM_002614 | SEQ ID NO 792 | Contig30480_RC | SEQ ID NO 2137 |
| | NM_002618 | SEQ ID NO 794 | Contig30609_RC | SEQ ID NO 2138 |
| | NM_002626 | SEQ ID NO 795 | Contig30934_RC | SEQ ID NO 2139 |
| 15 | NM_002633 | SEQ ID NO 796 | Contig31150_RC | SEQ ID NO 2140 |
| | NM_002639 | SEQ ID NO 797 | Contig31186_RC | SEQ ID NO 2141 |
| | NM_002648 | SEQ ID NO 798 | Contig31251_RC | SEQ ID NO 2142 |
| | NM_002659 | SEQ ID NO 799 | Contig31288_RC | SEQ ID NO 2143 |
| | NM_002661 | SEQ ID NO 800 | Contig31291_RC | SEQ ID NO 2144 |
| | NM_002662 | SEQ ID NO 801 | Contig31295_RC | SEQ ID NO 2145 |
| 20 | NM_002664 | SEQ ID NO 802 | Contig31424_RC | SEQ ID NO 2146 |
| | NM_002689 | SEQ ID NO 804 | Contig31449_RC | SEQ ID NO 2147 |
| | NM_002690 | SEQ ID NO 805 | Contig31596_RC | SEQ ID NO 2148 |
| | NM_002709 | SEQ ID NO 806 | Contig31864_RC | SEQ ID NO 2149 |
| | NM_002727 | SEQ ID NO 807 | Contig31928_RC | SEQ ID NO 2150 |
| 25 | NM_002729 | SEQ ID NO 808 | Contig31966_RC | SEQ ID NO 2151 |
| | NM_002734 | SEQ ID NO 809 | Contig31986_RC | SEQ ID NO 2152 |
| | NM_002736 | SEQ ID NO 810 | Contig32084_RC | SEQ ID NO 2153 |
| | NM_002740 | SEQ ID NO 811 | Contig32105_RC | SEQ ID NO 2154 |
| | NM_002748 | SEQ ID NO 813 | Contig32185_RC | SEQ ID NO 2156 |
| | NM_002774 | SEQ ID NO 814 | Contig32242_RC | SEQ ID NO 2157 |
| 30 | NM_002775 | SEQ ID NO 815 | Contig32322_RC | SEQ ID NO 2158 |
| | NM_002776 | SEQ ID NO 816 | Contig32336_RC | SEQ ID NO 2159 |
| | NM_002789 | SEQ ID NO 817 | Contig32558_RC | SEQ ID NO 2160 |
| | NM_002794 | SEQ ID NO 818 | Contig32798_RC | SEQ ID NO 2161 |
| | NM_002796 | SEQ ID NO 819 | Contig33005_RC | SEQ ID NO 2162 |
| 35 | NM_002800 | SEQ ID NO 820 | Contig33230_RC | SEQ ID NO 2163 |
| | NM_002801 | SEQ ID NO 821 | Contig33260_RC | SEQ ID NO 2164 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|---------------|-----------------------------|----------------|
| | NM_002808 | SEQ ID NO 822 | Contig33654_RC | SEQ ID NO 2166 |
| | NM_002821 | SEQ ID NO 824 | Contig33741_RC | SEQ ID NO 2167 |
| 5 | NM_002826 | SEQ ID NO 825 | Contig33771_RC | SEQ ID NO 2168 |
| | NM_002827 | SEQ ID NO 826 | Contig33814_RC | SEQ ID NO 2169 |
| | NM_002838 | SEQ ID NO 827 | Contig33815_RC | SEQ ID NO 2170 |
| | NM_002852 | SEQ ID NO 828 | Contig33833 | SEQ ID NO 2171 |
| | NM_002854 | SEQ ID NO 829 | Contig33998_RC | SEQ ID NO 2172 |
| | NM_002856 | SEQ ID NO 830 | Contig34079 | SEQ ID NO 2173 |
| 10 | NM_002857 | SEQ ID NO 831 | Contig34080_RC | SEQ ID NO 2174 |
| | NM_002858 | SEQ ID NO 832 | Contig34222_RC | SEQ ID NO 2175 |
| | NM_002888 | SEQ ID NO 833 | Contig34233_RC | SEQ ID NO 2176 |
| | NM_002890 | SEQ ID NO 834 | Contig34303_RC | SEQ ID NO 2177 |
| | NM_002901 | SEQ ID NO 836 | Contig34393_RC | SEQ ID NO 2178 |
| 15 | NM_002906 | SEQ ID NO 837 | Contig34477_RC | SEQ ID NO 2179 |
| | NM_002916 | SEQ ID NO 838 | Contig34766_RC | SEQ ID NO 2181 |
| | NM_002923 | SEQ ID NO 839 | Contig34952 | SEQ ID NO 2182 |
| | NM_002933 | SEQ ID NO 840 | Contig34989_RC | SEQ ID NO 2183 |
| | NM_002936 | SEQ ID NO 841 | Contig35030_RC | SEQ ID NO 2184 |
| | NM_002937 | SEQ ID NO 842 | Contig35251_RC | SEQ ID NO 2185 |
| 20 | NM_002950 | SEQ ID NO 843 | Contig35629_RC | SEQ ID NO 2186 |
| | NM_002961 | SEQ ID NO 844 | Contig35635_RC | SEQ ID NO 2187 |
| | NM_002964 | SEQ ID NO 845 | Contig35763_RC | SEQ ID NO 2188 |
| | NM_002965 | SEQ ID NO 846 | Contig35814_RC | SEQ ID NO 2189 |
| | NM_002966 | SEQ ID NO 847 | Contig35896_RC | SEQ ID NO 2190 |
| 25 | NM_002982 | SEQ ID NO 849 | Contig35976_RC | SEQ ID NO 2191 |
| | NM_002983 | SEQ ID NO 850 | Contig36042_RC | SEQ ID NO 2192 |
| | NM_002984 | SEQ ID NO 851 | Contig36081_RC | SEQ ID NO 2193 |
| | NM_002985 | SEQ ID NO 852 | Contig36152_RC | SEQ ID NO 2194 |
| | NM_002988 | SEQ ID NO 853 | Contig36193_RC | SEQ ID NO 2195 |
| | NM_002996 | SEQ ID NO 854 | Contig36312_RC | SEQ ID NO 2196 |
| 30 | NM_002997 | SEQ ID NO 855 | Contig36323_RC | SEQ ID NO 2197 |
| | NM_002999 | SEQ ID NO 856 | Contig36339_RC | SEQ ID NO 2198 |
| | NM_003012 | SEQ ID NO 857 | Contig36647_RC | SEQ ID NO 2199 |
| | NM_003022 | SEQ ID NO 858 | Contig36744_RC | SEQ ID NO 2200 |
| | NM_003034 | SEQ ID NO 859 | Contig36761_RC | SEQ ID NO 2201 |
| 35 | NM_003035 | SEQ ID NO 860 | Contig36879_RC | SEQ ID NO 2202 |
| | NM_003039 | SEQ ID NO 861 | Contig36900_RC | SEQ ID NO 2203 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|---------------|-----------------------------|----------------|
| | NM_003051 | SEQ ID NO 862 | Contig37015_RC | SEQ ID NO 2204 |
| | NM_003064 | SEQ ID NO 863 | Contig37024_RC | SEQ ID NO 2205 |
| 5 | NM_003066 | SEQ ID NO 864 | Contig37072_RC | SEQ ID NO 2207 |
| | NM_003088 | SEQ ID NO 865 | Contig37140_RC | SEQ ID NO 2208 |
| | NM_003090 | SEQ ID NO 866 | Contig37141_RC | SEQ ID NO 2209 |
| | NM_003096 | SEQ ID NO 867 | Contig37204_RC | SEQ ID NO 2210 |
| | NM_003099 | SEQ ID NO 868 | Contig37281_RC | SEQ ID NO 2211 |
| | NM_003102 | SEQ ID NO 869 | Contig37287_RC | SEQ ID NO 2212 |
| 10 | NM_003104 | SEQ ID NO 870 | Contig37439_RC | SEQ ID NO 2213 |
| | NM_003108 | SEQ ID NO 871 | Contig37562_RC | SEQ ID NO 2214 |
| | NM_003121 | SEQ ID NO 873 | Contig37571_RC | SEQ ID NO 2215 |
| | NM_003134 | SEQ ID NO 874 | Contig37598 | SEQ ID NO 2216 |
| | NM_003137 | SEQ ID NO 875 | Contig37758_RC | SEQ ID NO 2217 |
| 15 | NM_003144 | SEQ ID NO 876 | Contig37778_RC | SEQ ID NO 2218 |
| | NM_003146 | SEQ ID NO 877 | Contig37884_RC | SEQ ID NO 2219 |
| | NM_003149 | SEQ ID NO 878 | Contig37946_RC | SEQ ID NO 2220 |
| | NM_003151 | SEQ ID NO 879 | Contig38170_RC | SEQ ID NO 2221 |
| | NM_003157 | SEQ ID NO 880 | Contig38288_RC | SEQ ID NO 2223 |
| | NM_003158 | SEQ ID NO 881 | Contig38398_RC | SEQ ID NO 2224 |
| 20 | NM_003165 | SEQ ID NO 882 | Contig38580_RC | SEQ ID NO 2226 |
| | NM_003172 | SEQ ID NO 883 | Contig38630_RC | SEQ ID NO 2227 |
| | NM_003177 | SEQ ID NO 884 | Contig38652_RC | SEQ ID NO 2228 |
| | NM_003197 | SEQ ID NO 885 | Contig38683_RC | SEQ ID NO 2229 |
| | NM_003202 | SEQ ID NO 886 | Contig38726_RC | SEQ ID NO 2230 |
| 25 | NM_003213 | SEQ ID NO 887 | Contig38791_RC | SEQ ID NO 2231 |
| | NM_003217 | SEQ ID NO 888 | Contig38901_RC | SEQ ID NO 2232 |
| | NM_003225 | SEQ ID NO 889 | Contig38983_RC | SEQ ID NO 2233 |
| | NM_003226 | SEQ ID NO 890 | Contig39090_RC | SEQ ID NO 2234 |
| | NM_003236 | SEQ ID NO 892 | Contig39132_RC | SEQ ID NO 2235 |
| | NM_003239 | SEQ ID NO 893 | Contig39157_RC | SEQ ID NO 2236 |
| 30 | NM_003248 | SEQ ID NO 894 | Contig39226_RC | SEQ ID NO 2237 |
| | NM_003255 | SEQ ID NO 895 | Contig39285_RC | SEQ ID NO 2238 |
| | NM_003258 | SEQ ID NO 896 | Contig39556_RC | SEQ ID NO 2239 |
| | NM_003264 | SEQ ID NO 897 | Contig39591_RC | SEQ ID NO 2240 |
| | NM_003283 | SEQ ID NO 898 | Contig39826_RC | SEQ ID NO 2241 |
| 35 | NM_003318 | SEQ ID NO 899 | Contig39845_RC | SEQ ID NO 2242 |
| | NM_003329 | SEQ ID NO 900 | Contig39891_RC | SEQ ID NO 2243 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|---------------|-----------------------------|----------------|
| | NM_003332 | SEQ ID NO 901 | Contig39922_RC | SEQ ID NO 2244 |
| | NM_003358 | SEQ ID NO 902 | Contig39960_RC | SEQ ID NO 2245 |
| 5 | NM_003359 | SEQ ID NO 903 | Contig40026_RC | SEQ ID NO 2246 |
| | NM_003360 | SEQ ID NO 904 | Contig40121_RC | SEQ ID NO 2247 |
| | NM_003368 | SEQ ID NO 905 | Contig40128_RC | SEQ ID NO 2248 |
| | NM_003376 | SEQ ID NO 906 | Contig40146 | SEQ ID NO 2249 |
| | NM_003380 | SEQ ID NO 907 | Contig40208_RC | SEQ ID NO 2250 |
| | NM_003392 | SEQ ID NO 908 | Contig40212_RC | SEQ ID NO 2251 |
| 10 | NM_003412 | SEQ ID NO 909 | Contig40238_RC | SEQ ID NO 2252 |
| | NM_003430 | SEQ ID NO 910 | Contig40434_RC | SEQ ID NO 2253 |
| | NM_003462 | SEQ ID NO 911 | Contig40446_RC | SEQ ID NO 2254 |
| | NM_003467 | SEQ ID NO 912 | Contig40500_RC | SEQ ID NO 2255 |
| | NM_003472 | SEQ ID NO 913 | Contig40573_RC | SEQ ID NO 2256 |
| 15 | NM_003479 | SEQ ID NO 914 | Contig40813_RC | SEQ ID NO 2258 |
| | NM_003489 | SEQ ID NO 915 | Contig40816_RC | SEQ ID NO 2259 |
| | NM_003494 | SEQ ID NO 916 | Contig40845_RC | SEQ ID NO 2261 |
| | NM_003498 | SEQ ID NO 917 | Contig40889_RC | SEQ ID NO 2262 |
| | NM_003504 | SEQ ID NO 919 | Contig41035 | SEQ ID NO 2263 |
| | NM_003508 | SEQ ID NO 920 | Contig41234_RC | SEQ ID NO 2264 |
| 20 | NM_003510 | SEQ ID NO 921 | Contig41413_RC | SEQ ID NO 2266 |
| | NM_003512 | SEQ ID NO 922 | Contig41521_RC | SEQ ID NO 2267 |
| | NM_003528 | SEQ ID NO 923 | Contig41530_RC | SEQ ID NO 2268 |
| | NM_003544 | SEQ ID NO 924 | Contig41590 | SEQ ID NO 2269 |
| | NM_003561 | SEQ ID NO 925 | Contig41618_RC | SEQ ID NO 2270 |
| 25 | NM_003563 | SEQ ID NO 926 | Contig41624_RC | SEQ ID NO 2271 |
| | NM_003568 | SEQ ID NO 927 | Contig41635_RC | SEQ ID NO 2272 |
| | NM_003579 | SEQ ID NO 928 | Contig41676_RC | SEQ ID NO 2273 |
| | NM_003600 | SEQ ID NO 929 | Contig41689_RC | SEQ ID NO 2274 |
| | NM_003615 | SEQ ID NO 931 | Contig41804_RC | SEQ ID NO 2275 |
| | NM_003627 | SEQ ID NO 932 | Contig41887_RC | SEQ ID NO 2276 |
| 30 | NM_003645 | SEQ ID NO 935 | Contig41905_RC | SEQ ID NO 2277 |
| | NM_003651 | SEQ ID NO 936 | Contig41954_RC | SEQ ID NO 2278 |
| | NM_003657 | SEQ ID NO 937 | Contig41983_RC | SEQ ID NO 2279 |
| | NM_003662 | SEQ ID NO 938 | Contig42006_RC | SEQ ID NO 2280 |
| | NM_003670 | SEQ ID NO 939 | Contig42014_RC | SEQ ID NO 2281 |
| 35 | NM_003675 | SEQ ID NO 940 | Contig42036_RC | SEQ ID NO 2282 |
| | NM_003676 | SEQ ID NO 941 | Contig42041_RC | SEQ ID NO 2283 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|---------------|-----------------------------|----------------|
| | NM_003681 | SEQ ID NO 942 | Contig42139 | SEQ ID NO 2284 |
| | NM_003683 | SEQ ID NO 943 | Contig42161_RC | SEQ ID NO 2285 |
| 5 | NM_003686 | SEQ ID NO 944 | Contig42220_RC | SEQ ID NO 2286 |
| | NM_003689 | SEQ ID NO 945 | Contig42306_RC | SEQ ID NO 2287 |
| | NM_003714 | SEQ ID NO 946 | Contig42311_RC | SEQ ID NO 2288 |
| | NM_003720 | SEQ ID NO 947 | Contig42313_RC | SEQ ID NO 2289 |
| | NM_003726 | SEQ ID NO 948 | Contig42402_RC | SEQ ID NO 2290 |
| | NM_003729 | SEQ ID NO 949 | Contig42421_RC | SEQ ID NO 2291 |
| 10 | NM_003740 | SEQ ID NO 950 | Contig42430_RC | SEQ ID NO 2292 |
| | NM_003772 | SEQ ID NO 952 | Contig42431_RC | SEQ ID NO 2293 |
| | NM_003791 | SEQ ID NO 953 | Contig42542_RC | SEQ ID NO 2294 |
| | NM_003793 | SEQ ID NO 954 | Contig42582 | SEQ ID NO 2295 |
| | NM_003795 | SEQ ID NO 955 | Contig42631_RC | SEQ ID NO 2296 |
| 15 | NM_003806 | SEQ ID NO 956 | Contig42751_RC | SEQ ID NO 2297 |
| | NM_003821 | SEQ ID NO 957 | Contig42759_RC | SEQ ID NO 2298 |
| | NM_003829 | SEQ ID NO 958 | Contig43054 | SEQ ID NO 2299 |
| | NM_003831 | SEQ ID NO 959 | Contig43079_RC | SEQ ID NO 2300 |
| | NM_003862 | SEQ ID NO 960 | Contig43195_RC | SEQ ID NO 2301 |
| | NM_003866 | SEQ ID NO 961 | Contig43368_RC | SEQ ID NO 2302 |
| 20 | NM_003875 | SEQ ID NO 962 | Contig43410_RC | SEQ ID NO 2303 |
| | NM_003878 | SEQ ID NO 963 | Contig43476_RC | SEQ ID NO 2304 |
| | NM_003894 | SEQ ID NO 965 | Contig43549_RC | SEQ ID NO 2305 |
| | NM_003897 | SEQ ID NO 966 | Contig43645_RC | SEQ ID NO 2306 |
| | NM_003904 | SEQ ID NO 967 | Contig43648_RC | SEQ ID NO 2307 |
| 25 | NM_003929 | SEQ ID NO 968 | Contig43673_RC | SEQ ID NO 2308 |
| | NM_003933 | SEQ ID NO 969 | Contig43679_RC | SEQ ID NO 2309 |
| | NM_003937 | SEQ ID NO 970 | Contig43694_RC | SEQ ID NO 2310 |
| | NM_003940 | SEQ ID NO 971 | Contig43747_RC | SEQ ID NO 2311 |
| | NM_003942 | SEQ ID NO 972 | Contig43918_RC | SEQ ID NO 2312 |
| | NM_003944 | SEQ ID NO 973 | Contig43983_RC | SEQ ID NO 2313 |
| 30 | NM_003953 | SEQ ID NO 974 | Contig44040_RC | SEQ ID NO 2314 |
| | NM_003954 | SEQ ID NO 975 | Contig44064_RC | SEQ ID NO 2315 |
| | NM_003975 | SEQ ID NO 976 | Contig44195_RC | SEQ ID NO 2316 |
| | NM_003981 | SEQ ID NO 977 | Contig44226_RC | SEQ ID NO 2317 |
| | NM_003982 | SEQ ID NO 978 | Contig44289_RC | SEQ ID NO 2320 |
| 35 | NM_003986 | SEQ ID NO 979 | Contig44310_RC | SEQ ID NO 2321 |
| | NM_004003 | SEQ ID NO 980 | Contig44409 | SEQ ID NO 2322 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|----------------|-----------------------------|----------------|
| | NM_004010 | SEQ ID NO 981 | Contig44413_RC | SEQ ID NO 2323 |
| | NM_004024 | SEQ ID NO 982 | Contig44451_RC | SEQ ID NO 2324 |
| 5 | NM_004038 | SEQ ID NO 983 | Contig44585_RC | SEQ ID NO 2325 |
| | NM_004049 | SEQ ID NO 984 | Contig44656_RC | SEQ ID NO 2326 |
| | NM_004052 | SEQ ID NO 985 | Contig44703_RC | SEQ ID NO 2327 |
| | NM_004053 | SEQ ID NO 986 | Contig44708_RC | SEQ ID NO 2328 |
| | NM_004079 | SEQ ID NO 987 | Contig44757_RC | SEQ ID NO 2329 |
| | NM_004104 | SEQ ID NO 988 | Contig44829_RC | SEQ ID NO 2331 |
| 10 | NM_004109 | SEQ ID NO 989 | Contig44870 | SEQ ID NO 2332 |
| | NM_004110 | SEQ ID NO 990 | Contig44893_RC | SEQ ID NO 2333 |
| | NM_004120 | SEQ ID NO 991 | Contig44909_RC | SEQ ID NO 2334 |
| | NM_004131 | SEQ ID NO 992 | Contig44939_RC | SEQ ID NO 2335 |
| | NM_004143 | SEQ ID NO 993 | Contig45022_RC | SEQ ID NO 2336 |
| 15 | NM_004154 | SEQ ID NO 994 | Contig45032_RC | SEQ ID NO 2337 |
| | NM_004170 | SEQ ID NO 996 | Contig45041_RC | SEQ ID NO 2338 |
| | NM_004172 | SEQ ID NO 997 | Contig45049_RC | SEQ ID NO 2339 |
| | NM_004176 | SEQ ID NO 998 | Contig45090_RC | SEQ ID NO 2340 |
| | NM_004180 | SEQ ID NO 999 | Contig45156_RC | SEQ ID NO 2341 |
| | NM_004181 | SEQ ID NO 1000 | Contig45316_RC | SEQ ID NO 2342 |
| 20 | NM_004184 | SEQ ID NO 1001 | Contig45321 | SEQ ID NO 2343 |
| | NM_004203 | SEQ ID NO 1002 | Contig45375_RC | SEQ ID NO 2345 |
| | NM_004207 | SEQ ID NO 1003 | Contig45443_RC | SEQ ID NO 2346 |
| | NM_004217 | SEQ ID NO 1004 | Contig45454_RC | SEQ ID NO 2347 |
| | NM_004219 | SEQ ID NO 1005 | Contig45537_RC | SEQ ID NO 2348 |
| 25 | NM_004221 | SEQ ID NO 1006 | Contig45588_RC | SEQ ID NO 2349 |
| | NM_004233 | SEQ ID NO 1007 | Contig45708_RC | SEQ ID NO 2350 |
| | NM_004244 | SEQ ID NO 1008 | Contig45816_RC | SEQ ID NO 2351 |
| | NM_004252 | SEQ ID NO 1009 | Contig45847_RC | SEQ ID NO 2352 |
| | NM_004265 | SEQ ID NO 1010 | Contig45891_RC | SEQ ID NO 2353 |
| | NM_004267 | SEQ ID NO 1011 | Contig46056_RC | SEQ ID NO 2354 |
| 30 | NM_004281 | SEQ ID NO 1012 | Contig46062_RC | SEQ ID NO 2355 |
| | NM_004289 | SEQ ID NO 1013 | Contig46075_RC | SEQ ID NO 2356 |
| | NM_004298 | SEQ ID NO 1015 | Contig46164_RC | SEQ ID NO 2357 |
| | NM_004301 | SEQ ID NO 1016 | Contig46218_RC | SEQ ID NO 2358 |
| | NM_004305 | SEQ ID NO 1017 | Contig46223_RC | SEQ ID NO 2359 |
| 35 | NM_004311 | SEQ ID NO 1018 | Contig46244_RC | SEQ ID NO 2360 |
| | NM_004315 | SEQ ID NO 1019 | Contig46262_RC | SEQ ID NO 2361 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|----------------|-----------------------------|----------------|
| | NM_004323 | SEQ ID NO 1020 | Contig46362_RC | SEQ ID NO 2364 |
| | NM_004330 | SEQ ID NO 1021 | Contig46443_RC | SEQ ID NO 2365 |
| 5 | NM_004336 | SEQ ID NO 1022 | Contig46553_RC | SEQ ID NO 2367 |
| | NM_004338 | SEQ ID NO 1023 | Contig46597_RC | SEQ ID NO 2368 |
| | NM_004350 | SEQ ID NO 1024 | Contig46653_RC | SEQ ID NO 2369 |
| | NM_004354 | SEQ ID NO 1025 | Contig46709_RC | SEQ ID NO 2370 |
| | NM_004358 | SEQ ID NO 1026 | Contig46777_RC | SEQ ID NO 2371 |
| | NM_004360 | SEQ ID NO 1027 | Contig46802_RC | SEQ ID NO 2372 |
| 10 | NM_004362 | SEQ ID NO 1028 | Contig46890_RC | SEQ ID NO 2374 |
| | NM_004374 | SEQ ID NO 1029 | Contig46922_RC | SEQ ID NO 2375 |
| | NM_004378 | SEQ ID NO 1030 | Contig46934_RC | SEQ ID NO 2376 |
| | NM_004392 | SEQ ID NO 1031 | Contig46937_RC | SEQ ID NO 2377 |
| | NM_004395 | SEQ ID NO 1032 | Contig46991_RC | SEQ ID NO 2378 |
| 15 | NM_004414 | SEQ ID NO 1033 | Contig47016_RC | SEQ ID NO 2379 |
| | NM_004418 | SEQ ID NO 1034 | Contig47045_RC | SEQ ID NO 2380 |
| | NM_004425 | SEQ ID NO 1035 | Contig47106_RC | SEQ ID NO 2381 |
| | NM_004431 | SEQ ID NO 1036 | Contig47146_RC | SEQ ID NO 2382 |
| | NM_004436 | SEQ ID NO 1037 | Contig47230_RC | SEQ ID NO 2383 |
| | NM_004438 | SEQ ID NO 1038 | Contig47405_RC | SEQ ID NO 2384 |
| 20 | NM_004443 | SEQ ID NO 1039 | Contig47456_RC | SEQ ID NO 2385 |
| | NM_004446 | SEQ ID NO 1040 | Contig47465_RC | SEQ ID NO 2386 |
| | NM_004451 | SEQ ID NO 1041 | Contig47498_RC | SEQ ID NO 2387 |
| | NM_004454 | SEQ ID NO 1042 | Contig47578_RC | SEQ ID NO 2388 |
| | NM_004456 | SEQ ID NO 1043 | Contig47645_RC | SEQ ID NO 2389 |
| 25 | NM_004458 | SEQ ID NO 1044 | Contig47680_RC | SEQ ID NO 2390 |
| | NM_004472 | SEQ ID NO 1045 | Contig47781_RC | SEQ ID NO 2391 |
| | NM_004480 | SEQ ID NO 1046 | Contig47814_RC | SEQ ID NO 2392 |
| | NM_004482 | SEQ ID NO 1047 | Contig48004_RC | SEQ ID NO 2393 |
| | NM_004494 | SEQ ID NO 1048 | Contig48043_RC | SEQ ID NO 2394 |
| | NM_004496 | SEQ ID NO 1049 | Contig48057_RC | SEQ ID NO 2395 |
| 30 | NM_004503 | SEQ ID NO 1050 | Contig48076_RC | SEQ ID NO 2396 |
| | NM_004504 | SEQ ID NO 1051 | Contig48249_RC | SEQ ID NO 2397 |
| | NM_004515 | SEQ ID NO 1052 | Contig48263_RC | SEQ ID NO 2398 |
| | NM_004522 | SEQ ID NO 1053 | Contig48270_RC | SEQ ID NO 2399 |
| | NM_004523 | SEQ ID NO 1054 | Contig48328_RC | SEQ ID NO 2400 |
| 35 | NM_004525 | SEQ ID NO 1055 | Contig48518_RC | SEQ ID NO 2401 |
| | NM_004556 | SEQ ID NO 1056 | Contig48572_RC | SEQ ID NO 2402 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|----------------|-----------------------------|----------------|
| | NM_004559 | SEQ ID NO 1057 | Contig48659_RC | SEQ ID NO 2403 |
| | NM_004569 | SEQ ID NO 1058 | Contig48722_RC | SEQ ID NO 2404 |
| 5 | NM_004577 | SEQ ID NO 1059 | Contig48774_RC | SEQ ID NO 2405 |
| | NM_004585 | SEQ ID NO 1060 | Contig48776_RC | SEQ ID NO 2406 |
| | NM_004587 | SEQ ID NO 1061 | Contig48800_RC | SEQ ID NO 2407 |
| | NM_004594 | SEQ ID NO 1062 | Contig48806_RC | SEQ ID NO 2408 |
| | NM_004599 | SEQ ID NO 1063 | Contig48852_RC | SEQ ID NO 2409 |
| | NM_004633 | SEQ ID NO 1066 | Contig48900_RC | SEQ ID NO 2410 |
| 10 | NM_004642 | SEQ ID NO 1067 | Contig48913_RC | SEQ ID NO 2411 |
| | NM_004648 | SEQ ID NO 1068 | Contig48970_RC | SEQ ID NO 2413 |
| | NM_004663 | SEQ ID NO 1069 | Contig49058_RC | SEQ ID NO 2414 |
| | NM_004664 | SEQ ID NO 1070 | Contig49063_RC | SEQ ID NO 2415 |
| | NM_004684 | SEQ ID NO 1071 | Contig49093 | SEQ ID NO 2416 |
| 15 | NM_004688 | SEQ ID NO 1072 | Contig49098_RC | SEQ ID NO 2417 |
| | NM_004694 | SEQ ID NO 1073 | Contig49169_RC | SEQ ID NO 2418 |
| | NM_004695 | SEQ ID NO 1074 | Contig49233_RC | SEQ ID NO 2419 |
| | NM_004701 | SEQ ID NO 1075 | Contig49270_RC | SEQ ID NO 2420 |
| | NM_004708 | SEQ ID NO 1077 | Contig49282_RC | SEQ ID NO 2421 |
| | NM_004711 | SEQ ID NO 1078 | Contig49289_RC | SEQ ID NO 2422 |
| 20 | NM_004726 | SEQ ID NO 1079 | Contig49342_RC | SEQ ID NO 2423 |
| | NM_004750 | SEQ ID NO 1081 | Contig49344 | SEQ ID NO 2424 |
| | NM_004761 | SEQ ID NO 1082 | Contig49388_RC | SEQ ID NO 2425 |
| | NM_004762 | SEQ ID NO 1083 | Contig49405_RC | SEQ ID NO 2426 |
| | NM_004780 | SEQ ID NO 1085 | Contig49445_RC | SEQ ID NO 2427 |
| 25 | NM_004791 | SEQ ID NO 1086 | Contig49468_RC | SEQ ID NO 2428 |
| | NM_004798 | SEQ ID NO 1087 | Contig49509_RC | SEQ ID NO 2429 |
| | NM_004808 | SEQ ID NO 1088 | Contig49578_RC | SEQ ID NO 2431 |
| | NM_004811 | SEQ ID NO 1089 | Contig49581_RC | SEQ ID NO 2432 |
| | NM_004833 | SEQ ID NO 1090 | Contig49631_RC | SEQ ID NO 2433 |
| | NM_004835 | SEQ ID NO 1091 | Contig49673_RC | SEQ ID NO 2435 |
| 30 | NM_004843 | SEQ ID NO 1092 | Contig49743_RC | SEQ ID NO 2436 |
| | NM_004847 | SEQ ID NO 1093 | Contig49790_RC | SEQ ID NO 2437 |
| | NM_004848 | SEQ ID NO 1094 | Contig49818_RC | SEQ ID NO 2438 |
| | NM_004864 | SEQ ID NO 1095 | Contig49849_RC | SEQ ID NO 2439 |
| | NM_004865 | SEQ ID NO 1096 | Contig49855 | SEQ ID NO 2440 |
| 35 | NM_004866 | SEQ ID NO 1097 | Contig49910_RC | SEQ ID NO 2441 |
| | NM_004877 | SEQ ID NO 1098 | Contig49948_RC | SEQ ID NO 2442 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|----------------|-----------------------------|----------------|
| | NM_004900 | SEQ ID NO 1099 | Contig50004_RC | SEQ ID NO 2443 |
| | NM_004906 | SEQ ID NO 1100 | Contig50094 | SEQ ID NO 2444 |
| 5 | NM_004910 | SEQ ID NO 1101 | Contig50120_RC | SEQ ID NO 2446 |
| | NM_004918 | SEQ ID NO 1103 | Contig50153_RC | SEQ ID NO 2447 |
| | NM_004923 | SEQ ID NO 1104 | Contig50189_RC | SEQ ID NO 2448 |
| | NM_004938 | SEQ ID NO 1105 | Contig50276_RC | SEQ ID NO 2449 |
| | NM_004951 | SEQ ID NO 1106 | Contig50288_RC | SEQ ID NO 2450 |
| | NM_004968 | SEQ ID NO 1107 | Contig50297_RC | SEQ ID NO 2451 |
| 10 | NM_004994 | SEQ ID NO 1108 | Contig50391_RC | SEQ ID NO 2452 |
| | NM_004999 | SEQ ID NO 1109 | Contig50410 | SEQ ID NO 2453 |
| | NM_005001 | SEQ ID NO 1110 | Contig50523_RC | SEQ ID NO 2454 |
| | NM_005002 | SEQ ID NO 1111 | Contig50529 | SEQ ID NO 2455 |
| | NM_005012 | SEQ ID NO 1112 | Contig50588_RC | SEQ ID NO 2456 |
| 15 | NM_005032 | SEQ ID NO 1113 | Contig50592 | SEQ ID NO 2457 |
| | NM_005044 | SEQ ID NO 1114 | Contig50669_RC | SEQ ID NO 2458 |
| | NM_005046 | SEQ ID NO 1115 | Contig50719_RC | SEQ ID NO 2460 |
| | NM_005049 | SEQ ID NO 1116 | Contig50728_RC | SEQ ID NO 2461 |
| | NM_005067 | SEQ ID NO 1117 | Contig50731_RC | SEQ ID NO 2462 |
| | NM_005077 | SEQ ID NO 1118 | Contig50802_RC | SEQ ID NO 2463 |
| 20 | NM_005080 | SEQ ID NO 1119 | Contig50822_RC | SEQ ID NO 2464 |
| | NM_005084 | SEQ ID NO 1120 | Contig50850_RC | SEQ ID NO 2466 |
| | NM_005130 | SEQ ID NO 1122 | Contig50860_RC | SEQ ID NO 2467 |
| | NM_005139 | SEQ ID NO 1123 | Contig50913_RC | SEQ ID NO 2468 |
| | NM_005168 | SEQ ID NO 1125 | Contig50950_RC | SEQ ID NO 2469 |
| 25 | NM_005190 | SEQ ID NO 1126 | Contig51066_RC | SEQ ID NO 2470 |
| | NM_005196 | SEQ ID NO 1127 | Contig51105_RC | SEQ ID NO 2472 |
| | NM_005213 | SEQ ID NO 1128 | Contig51117_RC | SEQ ID NO 2473 |
| | NM_005218 | SEQ ID NO 1129 | Contig51196_RC | SEQ ID NO 2474 |
| | NM_005235 | SEQ ID NO 1130 | Contig51235_RC | SEQ ID NO 2475 |
| | NM_005245 | SEQ ID NO 1131 | Contig51254_RC | SEQ ID NO 2476 |
| 30 | NM_005249 | SEQ ID NO 1132 | Contig51352_RC | SEQ ID NO 2477 |
| | NM_005257 | SEQ ID NO 1133 | Contig51369_RC | SEQ ID NO 2478 |
| | NM_005264 | SEQ ID NO 1134 | Contig51392_RC | SEQ ID NO 2479 |
| | NM_005271 | SEQ ID NO 1135 | Contig51403_RC | SEQ ID NO 2480 |
| | NM_005314 | SEQ ID NO 1136 | Contig51685_RC | SEQ ID NO 2483 |
| 35 | NM_005321 | SEQ ID NO 1137 | Contig51726_RC | SEQ ID NO 2484 |
| | NM_005322 | SEQ ID NO 1138 | Contig51742_RC | SEQ ID NO 2485 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|----------------|-----------------------------|----------------|
| | NM_005325 | SEQ ID NO 1139 | Contig51749_RC | SEQ ID NO 2486 |
| | NM_005326 | SEQ ID NO 1140 | Contig51775_RC | SEQ ID NO 2487 |
| 5 | NM_005335 | SEQ ID NO 1141 | Contig51800 | SEQ ID NO 2488 |
| | NM_005337 | SEQ ID NO 1142 | Contig51809_RC | SEQ ID NO 2489 |
| | NM_005342 | SEQ ID NO 1143 | Contig51821_RC | SEQ ID NO 2490 |
| | NM_005345 | SEQ ID NO 1144 | Contig51888_RC | SEQ ID NO 2491 |
| | NM_005357 | SEQ ID NO 1145 | Contig51953_RC | SEQ ID NO 2493 |
| | NM_005375 | SEQ ID NO 1146 | Contig51967_RC | SEQ ID NO 2495 |
| 10 | NM_005391 | SEQ ID NO 1147 | Contig51981_RC | SEQ ID NO 2496 |
| | NM_005408 | SEQ ID NO 1148 | Contig51994_RC | SEQ ID NO 2497 |
| | NM_005409 | SEQ ID NO 1149 | Contig52082_RC | SEQ ID NO 2498 |
| | NM_005410 | SEQ ID NO 1150 | Contig52094_RC | SEQ ID NO 2499 |
| | NM_005426 | SEQ ID NO 1151 | Contig52320 | SEQ ID NO 2500 |
| 15 | NM_005433 | SEQ ID NO 1152 | Contig52398_RC | SEQ ID NO 2501 |
| | NM_005441 | SEQ ID NO 1153 | Contig52425_RC | SEQ ID NO 2503 |
| | NM_005443 | SEQ ID NO 1154 | Contig52482_RC | SEQ ID NO 2504 |
| | NM_005483 | SEQ ID NO 1155 | Contig52543_RC | SEQ ID NO 2505 |
| | NM_005486 | SEQ ID NO 1156 | Contig52553_RC | SEQ ID NO 2506 |
| | NM_005496 | SEQ ID NO 1157 | Contig52579_RC | SEQ ID NO 2507 |
| 20 | NM_005498 | SEQ ID NO 1158 | Contig52603_RC | SEQ ID NO 2508 |
| | NM_005499 | SEQ ID NO 1159 | Contig52639_RC | SEQ ID NO 2509 |
| | NM_005514 | SEQ ID NO 1160 | Contig52641_RC | SEQ ID NO 2510 |
| | NM_005531 | SEQ ID NO 1162 | Contig52684 | SEQ ID NO 2511 |
| | NM_005538 | SEQ ID NO 1163 | Contig52705_RC | SEQ ID NO 2512 |
| 25 | NM_005541 | SEQ ID NO 1164 | Contig52720_RC | SEQ ID NO 2513 |
| | NM_005544 | SEQ ID NO 1165 | Contig52722_RC | SEQ ID NO 2514 |
| | NM_005548 | SEQ ID NO 1166 | Contig52723_RC | SEQ ID NO 2515 |
| | NM_005554 | SEQ ID NO 1167 | Contig52740_RC | SEQ ID NO 2516 |
| | NM_005555 | SEQ ID NO 1168 | Contig52779_RC | SEQ ID NO 2517 |
| | NM_005556 | SEQ ID NO 1169 | Contig52957_RC | SEQ ID NO 2518 |
| 30 | NM_005557 | SEQ ID NO 1170 | Contig52994_RC | SEQ ID NO 2519 |
| | NM_005558 | SEQ ID NO 1171 | Contig53022_RC | SEQ ID NO 2520 |
| | NM_005562 | SEQ ID NO 1172 | Contig53038_RC | SEQ ID NO 2521 |
| | NM_005563 | SEQ ID NO 1173 | Contig53047_RC | SEQ ID NO 2522 |
| | NM_005565 | SEQ ID NO 1174 | Contig53130 | SEQ ID NO 2523 |
| 35 | NM_005566 | SEQ ID NO 1175 | Contig53183_RC | SEQ ID NO 2524 |
| | NM_005572 | SEQ ID NO 1176 | Contig53242_RC | SEQ ID NO 2526 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|----------------|-----------------------------|----------------|
| | NM_005582 | SEQ ID NO 1177 | Contig53248_RC | SEQ ID NO 2527 |
| | NM_005608 | SEQ ID NO 1178 | Contig53260_RC | SEQ ID NO 2528 |
| 5 | NM_005614 | SEQ ID NO 1179 | Contig53296_RC | SEQ ID NO 2531 |
| | NM_005617 | SEQ ID NO 1180 | Contig53307_RC | SEQ ID NO 2532 |
| | NM_005620 | SEQ ID NO 1181 | Contig53314_RC | SEQ ID NO 2533 |
| | NM_005625 | SEQ ID NO 1182 | Contig53401_RC | SEQ ID NO 2534 |
| | NM_005651 | SEQ ID NO 1183 | Contig53550_RC | SEQ ID NO 2535 |
| | NM_005658 | SEQ ID NO 1184 | Contig53551_RC | SEQ ID NO 2536 |
| 10 | NM_005659 | SEQ ID NO 1185 | Contig53598_RC | SEQ ID NO 2537 |
| | NM_005667 | SEQ ID NO 1186 | Contig53646_RC | SEQ ID NO 2538 |
| | NM_005686 | SEQ ID NO 1187 | Contig53658_RC | SEQ ID NO 2539 |
| | NM_005690 | SEQ ID NO 1188 | Contig53698_RC | SEQ ID NO 2540 |
| | NM_005720 | SEQ ID NO 1190 | Contig53719_RC | SEQ ID NO 2541 |
| 15 | NM_005727 | SEQ ID NO 1191 | Contig53742_RC | SEQ ID NO 2542 |
| | NM_005733 | SEQ ID NO 1192 | Contig53757_RC | SEQ ID NO 2543 |
| | NM_005737 | SEQ ID NO 1193 | Contig53870_RC | SEQ ID NO 2544 |
| | NM_005742 | SEQ ID NO 1194 | Contig53952_RC | SEQ ID NO 2546 |
| | NM_005746 | SEQ ID NO 1195 | Contig53962_RC | SEQ ID NO 2547 |
| | NM_005749 | SEQ ID NO 1196 | Contig53968_RC | SEQ ID NO 2548 |
| 20 | NM_005760 | SEQ ID NO 1197 | Contig54113_RC | SEQ ID NO 2549 |
| | NM_005764 | SEQ ID NO 1198 | Contig54142_RC | SEQ ID NO 2550 |
| | NM_005794 | SEQ ID NO 1199 | Contig54232_RC | SEQ ID NO 2551 |
| | NM_005796 | SEQ ID NO 1200 | Contig54242_RC | SEQ ID NO 2552 |
| | NM_005804 | SEQ ID NO 1201 | Contig54260_RC | SEQ ID NO 2553 |
| 25 | NM_005813 | SEQ ID NO 1202 | Contig54263_RC | SEQ ID NO 2554 |
| | NM_005824 | SEQ ID NO 1203 | Contig54295_RC | SEQ ID NO 2555 |
| | NM_005825 | SEQ ID NO 1204 | Contig54318_RC | SEQ ID NO 2556 |
| | NM_005849 | SEQ ID NO 1205 | Contig54325_RC | SEQ ID NO 2557 |
| | NM_005853 | SEQ ID NO 1206 | Contig54389_RC | SEQ ID NO 2558 |
| | NM_005855 | SEQ ID NO 1207 | Contig54394_RC | SEQ ID NO 2559 |
| 30 | NM_005864 | SEQ ID NO 1208 | Contig54414_RC | SEQ ID NO 2560 |
| | NM_005874 | SEQ ID NO 1209 | Contig54425 | SEQ ID NO 2561 |
| | NM_005876 | SEQ ID NO 1210 | Contig54477_RC | SEQ ID NO 2562 |
| | NM_005880 | SEQ ID NO 1211 | Contig54503_RC | SEQ ID NO 2563 |
| | NM_005891 | SEQ ID NO 1212 | Contig54534_RC | SEQ ID NO 2564 |
| 35 | NM_005892 | SEQ ID NO 1213 | Contig54560_RC | SEQ ID NO 2566 |
| | NM_005899 | SEQ ID NO 1214 | Contig54581_RC | SEQ ID NO 2567 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|----------------|-----------------------------|----------------|
| | NM_005915 | SEQ ID NO 1215 | Contig54609_RC | SEQ ID NO 2568 |
| | NM_005919 | SEQ ID NO 1216 | Contig54666_RC | SEQ ID NO 2569 |
| 5 | NM_005923 | SEQ ID NO 1217 | Contig54667_RC | SEQ ID NO 2570 |
| | NM_005928 | SEQ ID NO 1218 | Contig54726_RC | SEQ ID NO 2571 |
| | NM_005932 | SEQ ID NO 1219 | Contig54742_RC | SEQ ID NO 2572 |
| | NM_005935 | SEQ ID NO 1220 | Contig54745_RC | SEQ ID NO 2573 |
| | NM_005945 | SEQ ID NO 1221 | Contig54757_RC | SEQ ID NO 2574 |
| | NM_005953 | SEQ ID NO 1222 | Contig54761_RC | SEQ ID NO 2575 |
| 10 | NM_005978 | SEQ ID NO 1223 | Contig54813_RC | SEQ ID NO 2576 |
| | NM_005990 | SEQ ID NO 1224 | Contig54867_RC | SEQ ID NO 2577 |
| | NM_006002 | SEQ ID NO 1225 | Contig54895_RC | SEQ ID NO 2578 |
| | NM_006004 | SEQ ID NO 1226 | Contig54898_RC | SEQ ID NO 2579 |
| | NM_006005 | SEQ ID NO 1227 | Contig54913_RC | SEQ ID NO 2580 |
| 15 | NM_006006 | SEQ ID NO 1228 | Contig54965_RC | SEQ ID NO 2582 |
| | NM_006017 | SEQ ID NO 1229 | Contig54968_RC | SEQ ID NO 2583 |
| | NM_006018 | SEQ ID NO 1230 | Contig55069_RC | SEQ ID NO 2584 |
| | NM_006023 | SEQ ID NO 1231 | Contig55181_RC | SEQ ID NO 2585 |
| | NM_006027 | SEQ ID NO 1232 | Contig55188_RC | SEQ ID NO 2586 |
| | NM_006029 | SEQ ID NO 1233 | Contig55221_RC | SEQ ID NO 2587 |
| 20 | NM_006033 | SEQ ID NO 1234 | Contig55254_RC | SEQ ID NO 2588 |
| | NM_006051 | SEQ ID NO 1235 | Contig55265_RC | SEQ ID NO 2589 |
| | NM_006055 | SEQ ID NO 1236 | Contig55377_RC | SEQ ID NO 2591 |
| | NM_006074 | SEQ ID NO 1237 | Contig55397_RC | SEQ ID NO 2592 |
| | NM_006086 | SEQ ID NO 1238 | Contig55448_RC | SEQ ID NO 2593 |
| 25 | NM_006087 | SEQ ID NO 1239 | Contig55468_RC | SEQ ID NO 2594 |
| | NM_006096 | SEQ ID NO 1240 | Contig55500_RC | SEQ ID NO 2595 |
| | NM_006101 | SEQ ID NO 1241 | Contig55538_RC | SEQ ID NO 2596 |
| | NM_006103 | SEQ ID NO 1242 | Contig55558_RC | SEQ ID NO 2597 |
| | NM_006111 | SEQ ID NO 1243 | Contig55606_RC | SEQ ID NO 2598 |
| | NM_006113 | SEQ ID NO 1244 | Contig55674_RC | SEQ ID NO 2599 |
| 30 | NM_006115 | SEQ ID NO 1245 | Contig55725_RC | SEQ ID NO 2600 |
| | NM_006117 | SEQ ID NO 1246 | Contig55728_RC | SEQ ID NO 2601 |
| | NM_006142 | SEQ ID NO 1247 | Contig55756_RC | SEQ ID NO 2602 |
| | NM_006144 | SEQ ID NO 1248 | Contig55769_RC | SEQ ID NO 2603 |
| | NM_006148 | SEQ ID NO 1249 | Contig55771_RC | SEQ ID NO 2605 |
| 35 | NM_006153 | SEQ ID NO 1250 | Contig55813_RC | SEQ ID NO 2607 |
| | NM_006159 | SEQ ID NO 1251 | Contig55829_RC | SEQ ID NO 2608 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|----------------|-----------------------------|----------------|
| | NM_006170 | SEQ ID NO 1252 | Contig55852_RC | SEQ ID NO 2609 |
| | NM_006197 | SEQ ID NO 1253 | Contig55883_RC | SEQ ID NO 2610 |
| 5 | NM_006224 | SEQ ID NO 1255 | Contig55920_RC | SEQ ID NO 2611 |
| | NM_006227 | SEQ ID NO 1256 | Contig55940_RC | SEQ ID NO 2612 |
| | NM_006235 | SEQ ID NO 1257 | Contig55950_RC | SEQ ID NO 2613 |
| | NM_006243 | SEQ ID NO 1258 | Contig55991_RC | SEQ ID NO 2614 |
| | NM_006264 | SEQ ID NO 1259 | Contig55997_RC | SEQ ID NO 2615 |
| | NM_006271 | SEQ ID NO 1261 | Contig56023_RC | SEQ ID NO 2616 |
| 10 | NM_006274 | SEQ ID NO 1262 | Contig56030_RC | SEQ ID NO 2617 |
| | NM_006290 | SEQ ID NO 1265 | Contig56093_RC | SEQ ID NO 2618 |
| | NM_006291 | SEQ ID NO 1266 | Contig56205_RC | SEQ ID NO 2621 |
| | NM_006296 | SEQ ID NO 1267 | Contig56270_RC | SEQ ID NO 2622 |
| | NM_006304 | SEQ ID NO 1268 | Contig56276_RC | SEQ ID NO 2623 |
| 15 | NM_006314 | SEQ ID NO 1269 | Contig56291_RC | SEQ ID NO 2624 |
| | NM_006332 | SEQ ID NO 1270 | Contig56298_RC | SEQ ID NO 2625 |
| | NM_006357 | SEQ ID NO 1271 | Contig56307 | SEQ ID NO 2627 |
| | NM_006366 | SEQ ID NO 1272 | Contig56390_RC | SEQ ID NO 2628 |
| | NM_006372 | SEQ ID NO 1273 | Contig56434_RC | SEQ ID NO 2629 |
| | NM_006377 | SEQ ID NO 1274 | Contig56457_RC | SEQ ID NO 2630 |
| 20 | NM_006378 | SEQ ID NO 1275 | Contig56534_RC | SEQ ID NO 2631 |
| | NM_006383 | SEQ ID NO 1276 | Contig56670_RC | SEQ ID NO 2632 |
| | NM_006389 | SEQ ID NO 1277 | Contig56678_RC | SEQ ID NO 2633 |
| | NM_006393 | SEQ ID NO 1278 | Contig56742_RC | SEQ ID NO 2634 |
| | NM_006398 | SEQ ID NO 1279 | Contig56759_RC | SEQ ID NO 2635 |
| 25 | NM_006406 | SEQ ID NO 1280 | Contig56765_RC | SEQ ID NO 2636 |
| | NM_006408 | SEQ ID NO 1281 | Contig56843_RC | SEQ ID NO 2637 |
| | NM_006410 | SEQ ID NO 1282 | Contig57011_RC | SEQ ID NO 2638 |
| | NM_006414 | SEQ ID NO 1283 | Contig57023_RC | SEQ ID NO 2639 |
| | NM_006417 | SEQ ID NO 1284 | Contig57057_RC | SEQ ID NO 2640 |
| | NM_006430 | SEQ ID NO 1285 | Contig57076_RC | SEQ ID NO 2641 |
| 30 | NM_006460 | SEQ ID NO 1286 | Contig57081_RC | SEQ ID NO 2642 |
| | NM_006461 | SEQ ID NO 1287 | Contig57091_RC | SEQ ID NO 2643 |
| | NM_006469 | SEQ ID NO 1288 | Contig57138_RC | SEQ ID NO 2644 |
| | NM_006470 | SEQ ID NO 1289 | Contig57173_RC | SEQ ID NO 2645 |
| | NM_006491 | SEQ ID NO 1290 | Contig57230_RC | SEQ ID NO 2646 |
| 35 | NM_006495 | SEQ ID NO 1291 | Contig57258_RC | SEQ ID NO 2647 |
| | NM_006500 | SEQ ID NO 1292 | Contig57270_RC | SEQ ID NO 2648 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|----------------|-----------------------------|----------------|
| | NM_006509 | SEQ ID NO 1293 | Contig57272_RC | SEQ ID NO 2649 |
| | NM_006516 | SEQ ID NO 1294 | Contig57344_RC | SEQ ID NO 2650 |
| 5 | NM_006533 | SEQ ID NO 1295 | Contig57430_RC | SEQ ID NO 2651 |
| | NM_006551 | SEQ ID NO 1296 | Contig57458_RC | SEQ ID NO 2652 |
| | NM_006556 | SEQ ID NO 1297 | Contig57493_RC | SEQ ID NO 2653 |
| | NM_006558 | SEQ ID NO 1298 | Contig57584_RC | SEQ ID NO 2654 |
| | NM_006564 | SEQ ID NO 1299 | Contig57595 | SEQ ID NO 2655 |
| | NM_006573 | SEQ ID NO 1300 | Contig57602_RC | SEQ ID NO 2656 |
| 10 | NM_006607 | SEQ ID NO 1301 | Contig57609_RC | SEQ ID NO 2657 |
| | NM_006622 | SEQ ID NO 1302 | Contig57610_RC | SEQ ID NO 2658 |
| | NM_006623 | SEQ ID NO 1303 | Contig57644_RC | SEQ ID NO 2659 |
| | NM_006636 | SEQ ID NO 1304 | Contig57725_RC | SEQ ID NO 2660 |
| | NM_006670 | SEQ ID NO 1305 | Contig57739_RC | SEQ ID NO 2661 |
| 15 | NM_006681 | SEQ ID NO 1306 | Contig57825_RC | SEQ ID NO 2662 |
| | NM_006682 | SEQ ID NO 1307 | Contig57864_RC | SEQ ID NO 2663 |
| | NM_006696 | SEQ ID NO 1308 | Contig57940_RC | SEQ ID NO 2664 |
| | NM_006698 | SEQ ID NO 1309 | Contig58260_RC | SEQ ID NO 2665 |
| | NM_006705 | SEQ ID NO 1310 | Contig58272_RC | SEQ ID NO 2666 |
| | NM_006739 | SEQ ID NO 1311 | Contig58301_RC | SEQ ID NO 2667 |
| 20 | NM_006748 | SEQ ID NO 1312 | Contig58368_RC | SEQ ID NO 2668 |
| | NM_006759 | SEQ ID NO 1313 | Contig58471_RC | SEQ ID NO 2669 |
| | NM_006762 | SEQ ID NO 1314 | Contig58755_RC | SEQ ID NO 2671 |
| | NM_006763 | SEQ ID NO 1315 | Contig59120_RC | SEQ ID NO 2672 |
| | NM_006769 | SEQ ID NO 1316 | Contig60157_RC | SEQ ID NO 2673 |
| 25 | NM_006770 | SEQ ID NO 1317 | Contig60864_RC | SEQ ID NO 2676 |
| | NM_006780 | SEQ ID NO 1318 | Contig61254_RC | SEQ ID NO 2677 |
| | NM_006787 | SEQ ID NO 1319 | Contig61815 | SEQ ID NO 2678 |
| | NM_006806 | SEQ ID NO 1320 | Contig61975 | SEQ ID NO 2679 |
| | NM_006813 | SEQ ID NO 1321 | Contig62306 | SEQ ID NO 2680 |
| | NM_006825 | SEQ ID NO 1322 | Contig62568_RC | SEQ ID NO 2681 |
| 30 | NM_006826 | SEQ ID NO 1323 | Contig62922_RC | SEQ ID NO 2682 |
| | NM_006829 | SEQ ID NO 1324 | Contig62964_RC | SEQ ID NO 2683 |
| | NM_006834 | SEQ ID NO 1325 | Contig63520_RC | SEQ ID NO 2685 |
| | NM_006835 | SEQ ID NO 1326 | Contig63649_RC | SEQ ID NO 2686 |
| | NM_006840 | SEQ ID NO 1327 | Contig63683_RC | SEQ ID NO 2687 |
| 35 | NM_006845 | SEQ ID NO 1328 | Contig63748_RC | SEQ ID NO 2688 |
| | NM_006847 | SEQ ID NO 1329 | Contig64502 | SEQ ID NO 2689 |

| | GenBank | | GenBank | |
|----|------------------|----------------|------------------|----------------|
| | Accession Number | SEQ ID NO | Accession Number | SEQ ID NO |
| | NM_006851 | SEQ ID NO 1330 | Contig64688 | SEQ ID NO 2690 |
| | NM_006855 | SEQ ID NO 1331 | Contig64775_RC | SEQ ID NO 2691 |
| 5 | NM_006864 | SEQ ID NO 1332 | Contig65227 | SEQ ID NO 2692 |
| | NM_006868 | SEQ ID NO 1333 | Contig65663 | SEQ ID NO 2693 |
| | NM_006875 | SEQ ID NO 1334 | Contig65785_RC | SEQ ID NO 2694 |
| | NM_006889 | SEQ ID NO 1336 | Contig65900 | SEQ ID NO 2695 |
| | NM_006892 | SEQ ID NO 1337 | Contig66219_RC | SEQ ID NO 2696 |
| | NM_006912 | SEQ ID NO 1338 | Contig66705_RC | SEQ ID NO 2697 |
| 10 | NM_006931 | SEQ ID NO 1341 | Contig66759_RC | SEQ ID NO 2698 |
| | NM_006941 | SEQ ID NO 1342 | Contig67182_RC | SEQ ID NO 2699 |
| | NM_006943 | SEQ ID NO 1343 | | |

15

20

25

30

35

Table 2. 550 preferred ER status markers drawn from Table 1.

| Identifier | Correlation | Name | Description |
|-------------------|-------------|----------|---|
| NM_002051 | 0.763977 | GATA3 | GATA-binding protein 3 |
| AB020689 | 0.753592 | KIAA0882 | KIAA0882 protein |
| 5 NM_001218 | 0.753225 | CA12 | carbonic anhydrase XII |
| NM_000125 | 0.748421 | ESR1 | estrogen receptor 1 |
| Contig56678_RC | 0.747816 | | ESTs |
| NM_004496 | 0.729116 | HNF3A | hepatocyte nuclear factor 3, alpha |
| NM_017732 | 0.713398 | FLJ20262 | hypothetical protein FLJ20262 |
| 10 NM_006806 | -0.712678 | BTG3 | BTG family, member 3 |
| Contig56390_RC | 0.705940 | | ESTs |
| Contig37571_RC | 0.704468 | | ESTs |
| NM_004559 | -0.701617 | NSEP1 | nuclease sensitive element binding protein 1 |
| 15 Contig50153_RC | -0.696652 | | ESTs, Weakly similar to LKHU proteoglycan link protein precursor [H.sapiens] |
| NM_012155 | 0.694332 | EMAP-2 | microtubule-associated protein like echinoderm EMAP |
| Contig237_RC | 0.687485 | FLJ21127 | hypothetical protein FLJ21127 |
| NM_019063 | -0.686064 | C2ORF2 | chromosome 2 open reading frame 2 |
| 20 NM_012219 | -0.680900 | MRAS | muscle RAS oncogene homolog |
| NM_001982 | 0.676114 | ERBB3 | v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 3 |
| NM_006623 | -0.675090 | PHGDH | phosphoglycerate dehydrogenase |
| NM_000636 | -0.674282 | SOD2 | superoxide dismutase 2, mitochondrial |
| 25 NM_006017 | -0.670353 | PROML1 | prominin (mouse)-like 1 |
| Contig57940_RC | 0.667915 | MAP-1 | MAP-1 protein |
| Contig46934_RC | 0.666908 | | ESTs, Weakly similar to JE0350 Anterior gradient-2 [H.sapiens] |
| NM_005080 | 0.665772 | XBP1 | X-box binding protein 1 |
| 30 NM_014246 | 0.665725 | CELSR1 | cadherin, EGF LAG seven-pass G-type receptor 1, flamingo (Drosophila) homolog |

| Identifier | Correlation | Name | Description |
|----------------|-------------|----------|---|
| Contig54667_RC | -0.663727 | | Human DNA sequence from clone RP1-187J11 on chromosome 6q11.1-22.33. Contains the gene for a novel protein similar to S. pombe and S. cerevisiae predicted proteins, the gene for a novel protein similar to protein kinase C inhibitors, the 3' end of the gene for a novel protein similar to Drosophila L82 and predicted worm proteins, ESTs, STSs, GSSs and two putative CpG islands |
| Contig51994_RC | 0.663715 | | ESTs, Weakly similar to B0416.1 [C.elegans] |
| NM_016337 | 0.663006 | RNB6 | RNB6 |
| NM_015640 | -0.660165 | PAI-RBP1 | PAI-1 mRNA-binding protein |
| X07834 | -0.657798 | SOD2 | superoxide dismutase 2, mitochondrial |
| NM_012319 | 0.657666 | LIV-1 | LIV-1 protein, estrogen regulated |
| Contig41887_RC | 0.656042 | | ESTs, Weakly similar to Homolog of rat Zymogen granule membrane protein [H.sapiens] |
| NM_003462 | 0.655349 | P28 | dynein, axonemal, light intermediate polypeptide |
| Contig58301_RC | 0.654268 | | Homo sapiens mRNA; cDNA DKFZp667D095 (from clone DKFZp667D095) |
| NM_005375 | 0.653783 | MYB | v-myb avian myeloblastosis viral oncogene homolog |
| NM_017447 | -0.652445 | YG81 | hypothetical protein LOC54149 |
| Contig924_RC | -0.650658 | | ESTs |
| M55914 | -0.650181 | MPB1 | MYC promoter-binding protein 1 |
| NM_006004 | -0.649819 | UQCRH | ubiquinol-cytochrome c reductase hinge protein |
| NM_000964 | 0.649072 | RARA | retinoic acid receptor, alpha |
| NM_013301 | 0.647583 | HSU79303 | protein predicted by clone 23882 |
| AB023211 | -0.647403 | PDI2 | peptidyl arginine deiminase, type II |
| NM_016629 | -0.646412 | LOC51323 | hypothetical protein |
| K02403 | 0.645532 | C4A | complement component 4A |
| NM_016405 | -0.642201 | HSU93243 | Ubc6p homolog |
| Contig46597_RC | 0.641733 | | ESTs |
| Contig55377_RC | 0.640310 | | ESTs |
| NM_001207 | 0.637800 | BTF3 | basic transcription factor 3 |

| Identifier | Correlation | Name | Description |
|-------------------|-------------|---------------|---|
| NM_018166 | 0.636422 | FLJ10647 | hypothetical protein FLJ10647 |
| AL110202 | -0.635398 | | Homo sapiens mRNA; cDNA DKFZp586I2022 (from clone DKFZp586I2022) |
| 5 AL133105 | -0.635201 | DKFZp434F2322 | hypothetical protein DKFZp434F2322 |
| NM_016839 | -0.635169 | RBMS1 | RNA binding motif, single stranded interacting protein 1 |
| 10 Contig53130 | -0.634812 | | ESTs, Weakly similar to hyperpolarization-activated cyclic nucleotide-gated channel hHCN2 [H.sapiens] |
| NM_018014 | -0.634460 | BCL11A | B-cell CLL/lymphoma 11A (zinc finger protein) |
| NM_006769 | -0.632197 | LMO4 | LIM domain only 4 |
| U92544 | 0.631170 | JCL-1 | hepatocellular carcinoma associated protein; breast cancer associated gene 1 |
| 15 Contig49233_RC | -0.631047 | | Homo sapiens, Similar to nuclear receptor binding factor 2, clone IMAGE:3463191, mRNA, partial cds |
| AL133033 | 0.629690 | KIAA1025 | KIAA1025 protein |
| 20 AL049265 | 0.629414 | | Homo sapiens mRNA; cDNA DKFZp564F053 (from clone DKFZp564F053) |
| NM_018728 | 0.627989 | MYO5C | myosin 5C |
| NM_004780 | 0.627856 | TCEAL1 | transcription elongation factor A (SII)-like 1 |
| Contig760_RC | 0.627132 | | ESTs |
| 25 Contig399_RC | 0.626543 | FLJ12538 | hypothetical protein FLJ12538 similar to ras-related protein RAB17 |
| M83822 | 0.625092 | CDC4L | cell division cycle 4-like |
| NM_001255 | -0.625089 | CDC20 | CDC20 (cell division cycle 20, S. cerevisiae, homolog) |
| NM_006739 | -0.624903 | MCM5 | minichromosome maintenance deficient (S. cerevisiae) 5 (cell division cycle 46) |
| 30 NM_002888 | -0.624664 | RARRES1 | retinoic acid receptor responder (tazarotene induced) 1 |
| NM_003197 | 0.623850 | TCEB1L | transcription elongation factor B (SIII), polypeptide 1-like |
| NM_006787 | 0.623625 | JCL-1 | hepatocellular carcinoma associated protein; breast cancer associated gene 1 |
| 35 Contig49342_RC | 0.622179 | | ESTs |

| Identifier | Correlation | Name | Description |
|------------------|-------------|----------|---|
| AL133619 | 0.621719 | | Homo sapiens mRNA; cDNA DKFZp434E2321 (from clone DKFZp434E2321); partial cds |
| AL133622 | 0.621577 | KIAA0876 | KIAA0876 protein |
| 5 NM_004648 | -0.621532 | PTPNS1 | protein tyrosine phosphatase, non-receptor type substrate 1 |
| NM_001793 | -0.621530 | CDH3 | cadherin 3, type 1, P-cadherin (placental) |
| NM_003217 | 0.620915 | TEGT | testis enhanced gene transcript (BAX inhibitor 1) |
| 10 NM_001551 | 0.620832 | IGBP1 | immunoglobulin (CD79A) binding protein 1 |
| NM_002539 | -0.620683 | ODC1 | ornithine decarboxylase 1 |
| Contig55997_RC | -0.619932 | | ESTs |
| NM_000633 | 0.619547 | BCL2 | B-cell CLL/lymphoma 2 |
| NM_016267 | -0.619096 | TONDU | TONDU |
| 15 Contig3659_RC | 0.618048 | FLJ21174 | hypothetical protein FLJ21174 |
| NM_000191 | 0.617250 | HMGCL | 3-hydroxymethyl-3-methylglutaryl-Coenzyme A lyase (hydroxymethylglutaricaciduria) |
| NM_001267 | 0.616890 | CHAD | chondroadherin |
| Contig39090_RC | 0.616385 | | ESTs |
| 20 AF055270 | -0.616268 | HSSG1 | heat-shock suppressed protein 1 |
| Contig43054 | 0.616015 | FLJ21603 | hypothetical protein FLJ21603 |
| NM_001428 | -0.615855 | ENO1 | enolase 1, (alpha) |
| Contig51369_RC | 0.615466 | | ESTs |
| Contig36647_RC | 0.615310 | GFRA1 | GDNF family receptor alpha 1 |
| 25 NM_014096 | -0.614832 | PRO1659 | PRO1659 protein |
| NM_015937 | 0.614735 | LOC51604 | CGI-06 protein |
| Contig49790_RC | -0.614463 | | ESTs |
| NM_006759 | -0.614279 | UGP2 | UDP-glucose pyrophosphorylase 2 |
| Contig53598_RC | -0.613787 | FLJ11413 | hypothetical protein FLJ11413 |
| AF113132 | -0.613561 | PSA | phosphoserine aminotransferase |
| 30 AK000004 | 0.613001 | | Homo sapiens mRNA for FLJ00004 protein, partial cds |
| Contig52543_RC | 0.612960 | | Homo sapiens cDNA FLJ13945 fis, clone Y79AA1000969 |
| AB032966 | -0.611917 | KIAA1140 | KIAA1140 protein |
| AL080192 | 0.611544 | | Homo sapiens cDNA: FLJ21238 fis, clone COL01115 |
| 35 X56807 | -0.610654 | DSC2 | desmocollin 2 |

| Identifier | Correlation | Name | Description |
|----------------|-------------|----------|---|
| Contig30390_RC | 0.609614 | | ESTs |
| AL137362 | 0.609121 | FLJ22237 | hypothetical protein FLJ22237 |
| NM_014211 | -0.608585 | GABRP | gamma-aminobutyric acid (GABA) A receptor, pi |
| NM_006696 | 0.608474 | SMAP | thyroid hormone receptor coactivating protein |
| Contig45588_RC | -0.608273 | | Homo sapiens cDNA: FLJ22610 fis, clone HSI04930 |
| NM_003358 | 0.608244 | UGCG | UDP-glucose ceramide glucosyltransferase |
| NM_006153 | -0.608129 | NCK1 | NCK adaptor protein 1 |
| NM_001453 | -0.606939 | FOXC1 | forkhead box C1 |
| Contig54666_RC | 0.606475 | | oy65e02.x1 NCI_CGAP_CLL1 Homo sapiens cDNA clone IMAGE:1670714 3' similar to TR:Q29168 Q29168 UNKNOWN PROTEIN ;, mRNA sequence. |
| NM_005945 | -0.605945 | MPB1 | MYC promoter-binding protein 1 |
| Contig55725_RC | -0.605841 | | ESTs, Moderately similar to T50635 hypothetical protein DKFZp762L0311.1 [H.sapiens] |
| Contig37015_RC | -0.605780 | | ESTs, Weakly similar to UAS3_HUMAN UBASH3A PROTEIN [H.sapiens] |
| AL157480 | -0.604362 | SH3BP1 | SH3-domain binding protein 1 |
| NM_005325 | -0.604310 | H1F1 | H1 histone family, member 1 |
| NM_001446 | -0.604061 | FABP7 | fatty acid binding protein 7, brain |
| Contig263_RC | 0.603318 | | Homo sapiens cDNA: FLJ23000 fis, clone LNG00194 |
| Contig8347_RC | -0.603311 | | ESTs |
| NM_002988 | -0.603279 | SCYA18 | small inducible cytokine subfamily A (Cys-Cys), member 18, pulmonary and activation-regulated |
| AF111849 | 0.603157 | HELO1 | homolog of yeast long chain polyunsaturated fatty acid elongation enzyme 2 |
| NM_014700 | 0.603042 | KIAA0665 | KIAA0665 gene product |
| NM_001814 | -0.602988 | CTSC | cathepsin C |
| AF116682 | -0.602350 | PRO2013 | hypothetical protein PRO2013 |
| AB037836 | 0.602024 | KIAA1415 | KIAA1415 protein |
| AB002301 | 0.602005 | KIAA0303 | KIAA0303 protein |

| Identifier | Correlation | Name | Description |
|-------------------|-------------|---------------|---|
| NM_002996 | -0.601841 | SCYD1 | small inducible cytokine subfamily D (Cys-X3-Cys), member 1 (fractalkine, neurotactin) |
| 5 NM_018410 | -0.601765 | DKFZp762E1312 | hypothetical protein DKFZp762E1312 |
| Contig49581_RC | -0.601571 | KIAA1350 | KIAA1350 protein |
| NM_003088 | -0.601458 | SNL | singed (Drosophila)-like (sea urchin fascin homolog like) |
| 10 Contig47045_RC | 0.601088 | | ESTs, Weakly similar to DP1_HUMAN POLYPOSIS LOCUS PROTEIN 1 [H.sapiens] |
| NM_001806 | -0.600954 | CEBPG | CCAAT/enhancer binding protein (C/EBP), gamma |
| NM_004374 | 0.600766 | COX6C | cytochrome c oxidase subunit VIc |
| Contig52641_RC | 0.600132 | | ESTs, Weakly similar to CENB MOUSE MAJOR CENTROMERE AUTOANTIGEN B [M.musculus] |
| 15 NM_000100 | -0.600127 | CSTB | cystatin B (stefin B) |
| NM_002250 | -0.600004 | KCNN4 | potassium intermediate/small conductance calcium-activated channel, subfamily N, member 4 |
| AB033035 | -0.599423 | KIAA1209 | KIAA1209 protein |
| Contig53968_RC | 0.599077 | | ESTs |
| 20 NM_002300 | -0.598246 | LDHB | lactate dehydrogenase B |
| NM_000507 | 0.598110 | FBP1 | fructose-1,6-bisphosphatase 1 |
| NM_002053 | -0.597756 | GBP1 | guanylate binding protein 1, interferon-inducible, 67kD |
| AB007883 | 0.597043 | KIAA0423 | KIAA0423 protein |
| 25 NM_004900 | -0.597010 | DJ742C19.2 | phorbolin (similar to apolipoprotein B mRNA editing protein) |
| NM_004480 | 0.596321 | FUT8 | fucosyltransferase 8 (alpha (1,6) fucosyltransferase) |
| Contig35896_RC | 0.596281 | | ESTs |
| NM_020974 | 0.595173 | CEGP1 | CEGP1 protein |
| 30 NM_000662 | 0.595114 | NAT1 | N-acetyltransferase 1 (arylamine N-acetyltransferase) |
| NM_006113 | 0.595017 | VAV3 | vav 3 oncogene |
| NM_014865 | -0.594928 | KIAA0159 | chromosome condensation-related SMC-associated protein 1 |
| Contig55538_RC | -0.594573 | BA395L14.2 | hypothetical protein bA395L14.2 |
| 35 NM_016056 | 0.594084 | LOC51643 | CGI-119 protein |

| Identifier | Correlation | Name | Description |
|-------------------|-------------|--------------|--|
| NM_003579 | -0.594063 | RAD54L | RAD54 (<i>S.cerevisiae</i>)-like |
| NM_014214 | -0.593860 | IMPA2 | inositol(myo)-1(or 4)-monophosphatase 2 |
| 5 U79293 | 0.593793 | | Human clone 23948 mRNA sequence |
| NM_005557 | -0.593746 | KRT16 | keratin 16 (focal non-epidermolytic palmoplantar keratoderma) |
| NM_002444 | -0.592405 | MSN | moesin |
| NM_003681 | -0.592155 | PDXK | pyridoxal (pyridoxine, vitamin B6) kinase |
| 10 NM_006372 | -0.591711 | NSAP1 | NS1-associated protein 1 |
| NM_005218 | -0.591192 | DEFB1 | defensin, beta 1 |
| NM_004642 | -0.591081 | DOC1 | deleted in oral cancer (mouse, homolog) 1 |
| 15 AL133074 | 0.590359 | | Homo sapiens cDNA: FLJ22139 fis, clone HEP20959 |
| M73547 | 0.590317 | D5S346 | DNA segment, single copy probe LNS-CAI/LNS-CAII (deleted in polyposis) |
| Contig65663 | 0.590312 | | ESTs |
| AL035297 | -0.589728 | | H.sapiens gene from PAC 747L4 |
| 20 Contig35629_RC | 0.589383 | | ESTs |
| NM_019027 | 0.588862 | FLJ20273 | hypothetical protein |
| NM_012425 | -0.588804 | | Homo sapiens Ras suppressor protein 1 (RSU1), mRNA |
| NM_020179 | -0.588326 | FN5 | FN5 protein |
| AF090913 | -0.587275 | TMSB10 | thymosin, beta 10 |
| 25 NM_004176 | 0.587190 | SREBF1 | sterol regulatory element binding transcription factor 1 |
| NM_016121 | 0.586941 | LOC51133 | NY-REN-45 antigen |
| NM_014773 | 0.586871 | KIAA0141 | KIAA0141 gene product |
| NM_019000 | 0.586677 | FLJ20152 | hypothetical protein |
| NM_016243 | 0.585942 | LOC51706 | cytochrome b5 reductase 1 (B5R.1) |
| 30 NM_014274 | -0.585815 | ABP/ZF | Alu-binding protein with zinc finger domain |
| NM_018379 | 0.585497 | FLJ11280 | hypothetical protein FLJ11280 |
| AL157431 | -0.585077 | DKFZp762A227 | hypothetical protein DKFZp762A227 |
| D38521 | -0.584684 | KIAA0077 | KIAA0077 protein |
| 35 NM_002570 | 0.584272 | PACE4 | paired basic amino acid cleaving system 4 |

| Identifier | Correlation | Name | Description |
|-------------------|-------------|----------|---|
| NM_001809 | -0.584252 | CENPA | centromere protein A (17kD) |
| NM_003318 | -0.583556 | TTK | TTK protein kinase |
| NM_014325 | -0.583555 | CORO1C | coronin, actin-binding protein, 1C |
| 5 NM_005667 | 0.583376 | ZFP103 | zinc finger protein homologous to Zfp103 in mouse |
| NM_004354 | 0.582420 | CCNG2 | cyclin G2 |
| NM_003670 | 0.582235 | BHLHB2 | basic helix-loop-helix domain containing, class B, 2 |
| NM_001673 | -0.581902 | ASNS | asparagine synthetase |
| 10 NM_001333 | -0.581402 | CTSL2 | cathepsin L2 |
| Contig54295_RC | 0.581256 | | ESTs |
| Contig33998_RC | 0.581018 | | ESTs |
| NM_006002 | -0.580592 | UCHL3 | ubiquitin carboxyl-terminal esterase L3 (ubiquitin thiolesterase) |
| 15 NM_015392 | 0.580568 | NPDC1 | neural proliferation, differentiation and control, 1 |
| NM_004866 | 0.580138 | SCAMP1 | secretory carrier membrane protein 1 |
| Contig50391_RC | 0.580071 | | ESTs |
| NM_000592 | 0.579965 | C4B | complement component 4B |
| Contig50802_RC | 0.579881 | | ESTs |
| 20 Contig41635_RC | -0.579468 | | ESTs |
| NM_006845 | -0.579339 | KNSL6 | kinesin-like 6 (mitotic centromere-associated kinesin) |
| NM_003720 | -0.579296 | DSCR2 | Down syndrome critical region gene 2 |
| NM_000060 | 0.578967 | BTB | biotinidase |
| 25 AL050388 | -0.578736 | | Homo sapiens mRNA; cDNA DKFZp564M2422 (from clone DKFZp564M2422); partial cds |
| NM_003772 | -0.578395 | JRKL | jerky (mouse) homolog-like |
| NM_014398 | -0.578388 | TSC403 | similar to lysosome-associated membrane glycoprotein |
| 30 NM_001280 | 0.578213 | CIRBP | cold inducible RNA-binding protein |
| NM_001395 | -0.577369 | DUSP9 | dual specificity phosphatase 9 |
| NM_016229 | -0.576290 | LOC51700 | cytochrome b5 reductase b5R.2 |
| NM_006096 | -0.575615 | NDRG1 | N-myc downstream regulated |
| NM_001552 | 0.575438 | IGFBP4 | insulin-like growth factor-binding protein 4 |
| 35 NM_005558 | -0.574818 | LAD1 | ladinin 1 |

| Identifier | Correlation | Name | Description |
|----------------|-------------|----------|--|
| Contig54534_RC | 0.574784 | | Human glucose transporter pseudogene |
| Contig1239_RC | 0.573822 | | Human Chromosome 16 BAC clone CIT987SK-A-362G6 |
| Contig57173_RC | 0.573807 | | Homo sapiens mRNA for KIAA1737 protein, partial cds |
| NM_004414 | -0.573538 | DSCR1 | Down syndrome critical region gene 1 |
| NM_021103 | -0.572722 | TMSB10 | thymosin, beta 10 |
| NM_002350 | -0.571917 | LYN | v-yes-1 Yamaguchi sarcoma viral related oncogene homolog |
| Contig51235_RC | 0.571049 | | Homo sapiens cDNA: FLJ23388 fis, clone HEP17008 |
| NM_013384 | 0.570987 | TMSG1 | tumor metastasis-suppressor |
| NM_014399 | 0.570936 | NET-6 | tetraspan NET-6 protein |
| Contig26022_RC | -0.570851 | | ESTs |
| AB023152 | 0.570561 | KIAA0935 | KIAA0935 protein |
| NM_021077 | -0.569944 | NMB | neuromedin B |
| NM_003498 | -0.569129 | SNN | stannin |
| U17077 | -0.568979 | BENE | BENE protein |
| D86985 | 0.567698 | KIAA0232 | KIAA0232 gene product |
| NM_006357 | -0.567513 | UBE2E3 | ubiquitin-conjugating enzyme E2E 3 (homologous to yeast UBC4/5) |
| AL049397 | -0.567434 | | Homo sapiens mRNA; cDNA DKFZp586C1019 (from clone DKFZp586C1019) |
| Contig64502 | 0.567433 | | ESTs, Weakly similar to unknown [M.musculus] |
| Contig56298_RC | -0.566892 | FLJ13154 | hypothetical protein FLJ13154 |
| Contig46056_RC | 0.566634 | | ESTs, Weakly similar to YZ28_HUMAN HYPOTHETICAL PROTEIN ZAP128 [H.sapiens] |
| AF007153 | 0.566044 | | Homo sapiens clone 23736 mRNA sequence |
| Contig1778_RC | -0.565789 | | ESTs |
| NM_017702 | -0.565789 | FLJ20186 | hypothetical protein FLJ20186 |
| Contig39226_RC | 0.565761 | | Homo sapiens cDNA FLJ12187 fis, clone MAMMA1000831 |
| NM_000168 | 0.564879 | GLI3 | GLI-Kruppel family member GLI3 (Greig cephalopolysyndactyly syndrome) |

| Identifier | Correlation | Name | Description |
|-------------------|-------------|----------|--|
| Contig57609_RC | 0.564751 | | ESTs, Weakly similar to T2D3 HUMAN TRANSCRIPTION INITIATION FACTOR TFIID 135 KDA SUBUNIT [H.sapiens] |
| 5 U45975 | 0.564602 | PIB5PA | phosphatidylinositol (4,5) biphosphate 5-phosphatase, A |
| AF038182 | 0.564596 | | Homo sapiens clone 23860 mRNA sequence |
| Contig5348_RC | 0.564480 | | ESTs, Weakly similar to 1607338A transcription factor BTF3a [H.sapiens] |
| 10 NM_001321 | -0.564459 | CSRP2 | cysteine and glycine-rich protein 2 |
| Contig25362_RC | -0.563801 | | ESTs |
| NM_001609 | 0.563782 | ACADSB | acyl-Coenzyme A dehydrogenase, short/branched chain |
| 15 Contig40146 | 0.563731 | | wi84e12.x1 NCI_CGAP_Kid12 Homo sapiens cDNA clone IMAGE:2400046 3' similar to SW:RASD DICDI P03967 RAS-LIKE PROTEIN RASD ;, mRNA sequence. |
| NM_016002 | 0.563403 | LOC51097 | CGI-49 protein |
| Contig34303_RC | 0.563157 | | Homo sapiens cDNA: FLJ21517 fis, clone COL05829 |
| 20 Contig55883_RC | 0.563141 | | ESTs |
| NM_017961 | 0.562479 | FLJ20813 | hypothetical protein FLJ20813 |
| M21551 | -0.562340 | NMB | neuromedin B |
| 25 Contig3940_RC | -0.561956 | YWHAH | tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, eta polypeptide |
| AB033111 | -0.561746 | KIAA1285 | KIAA1285 protein |
| Contig43410_RC | 0.561678 | | ESTs |
| Contig42006_RC | -0.561677 | | ESTs |
| Contig57272_RC | 0.561228 | | ESTs |
| 30 G26403 | -0.561068 | YWHAH | tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, eta polypeptide |
| NM_005915 | -0.560813 | MCM6 | minichromosome maintenance deficient (mis5, S. pombe) 6 |
| NM_003875 | -0.560668 | GMPS | guanine monphosphate synthetase |
| 35 AK000142 | 0.559651 | AK000142 | Homo sapiens cDNA FLJ20135 fis, clone COL06818. |

| Identifier | Correlation | Name | Description |
|------------------|-------------|--------------|--|
| NM_002709 | -0.559621 | PPP1CB | protein phosphatase 1, catalytic subunit, beta isoform |
| NM_001276 | -0.558868 | CHI3L1 | chitinase 3-like 1 (cartilage glycoprotein-39) |
| 5 NM_002857 | 0.558862 | PXF | peroxisomal farnesylated protein |
| Contig33815_RC | -0.558741 | FLJ22833 | hypothetical protein FLJ22833 |
| NM_003740 | -0.558491 | KCNK5 | potassium channel, subfamily K, member 5 (TASK-2) |
| Contig53646_RC | 0.558455 | | ESTs |
| 10 NM_005538 | -0.558350 | INHBC | inhibin, beta C |
| NM_002111 | 0.557860 | HD | huntingtin (Huntington disease) |
| NM_003683 | -0.557807 | D21S2056 E | DNA segment on chromosome 21 (unique) 2056 expressed sequence |
| NM_003035 | -0.557380 | SIL | TAL1 (SCL) interrupting locus |
| 15 Contig4388_RC | -0.557216 | | Homo sapiens, Similar to integral membrane protein 3, clone MGC:3011, mRNA, complete cds |
| Contig38288_RC | -0.556426 | | ESTs, Weakly similar to ISHUS protein disulfide-isomerase [H.sapiens] |
| NM_015417 | 0.556184 | DKFZP434I114 | DKFZP434I114 protein |
| 20 NM_015507 | -0.556138 | EGFL6 | EGF-like-domain, multiple 6 |
| AF279865 | 0.555951 | KIF13B | kinesin family member 13B |
| Contig31288_RC | -0.555754 | | ESTs |
| NM_002966 | -0.555620 | S100A10 | S100 calcium-binding protein A10 (annexin II ligand, calpactin I, light polypeptide (p11)) |
| 25 NM_017585 | -0.555476 | SLC2A6 | solute carrier family 2 (facilitated glucose transporter), member 6 |
| NM_013296 | -0.555367 | HSU54999 | LGN protein |
| NM_000224 | 0.554838 | KRT18 | keratin 18 |
| Contig49270_RC | -0.554593 | KIAA1553 | KIAA1553 protein |
| NM_004848 | -0.554538 | ICB-1 | basement membrane-induced gene |
| 30 NM_007275 | 0.554278 | FUS1 | lung cancer candidate |
| NM_007044 | -0.553550 | KATNA1 | katanin p60 (ATPase-containing) subunit A 1 |
| Contig1829 | 0.553317 | | ESTs |
| AF272357 | 0.553286 | NPDC1 | neural proliferation, differentiation and control, 1 |

35

| Identifier | Correlation | Name | Description |
|----------------|-------------|---------------|---|
| Contig57584_RC | -0.553080 | | Homo sapiens, Similar to gene rich cluster, C8 gene, clone MGC:2577, mRNA, complete cds |
| 5 NM_003039 | -0.552747 | SLC2A5 | solute carrier family 2 (facilitated glucose transporter), member 5 |
| NM_014216 | 0.552321 | ITPK1 | inositol 1,3,4-triphosphate 5/6 kinase |
| NM_007027 | -0.552064 | TOPBP1 | topoisomerase (DNA) II binding protein |
| 10 AF118224 | -0.551916 | ST14 | suppression of tumorigenicity 14 (colon carcinoma, matriptase, epithin) |
| X75315 | -0.551853 | HSRNASE B | seb4D |
| NM_012101 | -0.551824 | ATDC | ataxia-telangiectasia group D-associated protein |
| AL157482 | -0.551329 | FLJ23399 | hypothetical protein FLJ23399 |
| 15 NM_012474 | -0.551150 | UMPK | uridine monophosphate kinase |
| Contig57081_RC | 0.551103 | | ESTs |
| NM_006941 | -0.551069 | SOX10 | SRY (sex determining region Y)-box 10 |
| 20 NM_004694 | 0.550932 | SLC16A6 | solute carrier family 16 (monocarboxylic acid transporters), member 6 |
| Contig9541_RC | 0.550680 | | ESTs |
| Contig20617_RC | 0.550546 | | ESTs |
| NM_004252 | 0.550365 | SLC9A3R 1 | solute carrier family 9 (sodium/hydrogen exchanger), isoform 3 regulatory factor 1 |
| 25 NM_015641 | -0.550200 | DKFZP586B2022 | testin |
| NM_004336 | -0.550164 | BUB1 | budding uninhibited by benzimidazoles 1 (yeast homolog) |
| Contig39960_RC | -0.549951 | FLJ21079 | hypothetical protein FLJ21079 |
| NM_020686 | 0.549659 | NPD009 | NPD009 protein |
| 30 NM_002633 | -0.549647 | PGM1 | phosphoglucomutase 1 |
| Contig30480_RC | 0.548932 | | ESTs |
| NM_003479 | 0.548896 | PTP4A2 | protein tyrosine phosphatase type IVA, member 2 |
| NM_001679 | -0.548768 | ATP1B3 | ATPase, Na ⁺ /K ⁺ transporting, beta 3 polypeptide |
| 35 NM_001124 | -0.548601 | ADM | adrenomedullin |
| NM_001216 | -0.548375 | CA9 | carbonic anhydrase IX |

| Identifier | Correlation | Name | Description |
|------------------|-------------|---------------|---|
| U58033 | -0.548354 | MTMR2 | myotubularin related protein 2 |
| NM_018389 | -0.547875 | FLJ11320 | hypothetical protein FLJ11320 |
| AF176012 | 0.547867 | JDP1 | J domain containing protein 1 |
| 5 Contig66705_RC | -0.546926 | ST5 | suppression of tumorigenicity 5 |
| NM_018194 | 0.546878 | FLJ10724 | hypothetical protein FLJ10724 |
| NM_006851 | -0.546823 | RTVP1 | glioma pathogenesis-related protein |
| Contig53870_RC | 0.546756 | | ESTs |
| NM_002482 | -0.546012 | NASP | nuclear autoantigenic sperm protein (histone-binding) |
| 10 NM_002292 | 0.545949 | LAMB2 | laminin, beta 2 (laminin S) |
| NM_014696 | -0.545758 | KIAA0514 | KIAA0514 gene product |
| Contig49855 | 0.545517 | | ESTs |
| AL117666 | 0.545203 | DKFZP586O1624 | DKFZP586O1624 protein |
| NM_004701 | -0.545185 | CCNB2 | cyclin B2 |
| 15 NM_007050 | 0.544890 | PTPRT | protein tyrosine phosphatase, receptor type, T |
| NM_000414 | 0.544778 | HSD17B4 | hydroxysteroid (17-beta) dehydrogenase 4 |
| Contig52398_RC | -0.544775 | | Homo sapiens cDNA: FLJ21950 fis, clone HEP04949 |
| 20 AB007916 | 0.544496 | KIAA0447 | KIAA0447 gene product |
| Contig66219_RC | 0.544467 | FLJ22402 | hypothetical protein FLJ22402 |
| D87453 | 0.544145 | KIAA0264 | KIAA0264 protein |
| NM_015515 | -0.543929 | DKFZP434G032 | DKFZP434G032 protein |
| 25 NM_001530 | -0.543898 | HIF1A | hypoxia-inducible factor 1, alpha subunit (basic helix-loop-helix transcription factor) |
| NM_004109 | -0.543893 | FDX1 | ferredoxin 1 |
| NM_000381 | -0.543871 | MID1 | midline 1 (Opitz/BBB syndrome) |
| Contig43983_RC | 0.543523 | CS2 | calsyntenin-2 |
| 30 AL137761 | 0.543371 | | Homo sapiens mRNA; cDNA DKFZp586L2424 (from clone DKFZp586L2424) |
| NM_005764 | -0.543175 | DD96 | epithelial protein up-regulated in carcinoma, membrane associated protein 17 |
| Contig1838_RC | 0.542996 | | Homo sapiens cDNA: FLJ22722 fis, clone HSI14444 |
| 35 NM_006670 | 0.542932 | 5T4 | 5T4 oncofetal trophoblast glycoprotein |

| Identifier | Correlation | Name | Description |
|-------------------|-------------|----------|---|
| Contig28552_RC | -0.542617 | | Homo sapiens mRNA; cDNA DKFZp434C0931 (from clone DKFZp434C0931); partial cds |
| Contig14284_RC | 0.542224 | | ESTs |
| 5 NM_006290 | -0.542115 | TNFAIP3 | tumor necrosis factor, alpha-induced protein 3 |
| AL050372 | 0.541463 | | Homo sapiens mRNA; cDNA DKFZp434A091 (from clone DKFZp434A091); partial cds |
| NM_014181 | -0.541095 | HSPC159 | HSPC159 protein |
| 10 Contig37141_RC | 0.540990 | | Homo sapiens cDNA: FLJ23582 fis, clone LNG13759 |
| NM_000947 | -0.540621 | PRIM2A | primase, polypeptide 2A (58kD) |
| NM_002136 | 0.540572 | HNRPA1 | heterogeneous nuclear ribonucleoprotein A1 |
| NM_004494 | -0.540543 | HDGF | hepatoma-derived growth factor (high-mobility group protein 1-like) |
| 15 Contig38983_RC | 0.540526 | | ESTs |
| Contig27882_RC | -0.540506 | | ESTs |
| Z11887 | -0.540020 | MMP7 | matrix metalloproteinase 7 (matrilysin, uterine) |
| NM_014575 | -0.539725 | SCHIP-1 | schwannomin interacting protein 1 |
| 20 Contig38170_RC | 0.539708 | | ESTs |
| Contig44064_RC | 0.539403 | | ESTs |
| U68385 | 0.539395 | MEIS3 | Meis (mouse) homolog 3 |
| Contig51967_RC | 0.538952 | | ESTs |
| Contig37562_RC | 0.538657 | | ESTs, Weakly similar to transformation-related protein [H.sapiens] |
| 25 Contig40500_RC | 0.538582 | | ESTs, Weakly similar to unnamed protein product [H.sapiens] |
| Contig1129_RC | 0.538339 | | ESTs |
| NM_002184 | 0.538185 | IL6ST | interleukin 6 signal transducer (gp130, oncostatin M receptor) |
| 30 AL049381 | 0.538041 | | Homo sapiens cDNA FLJ12900 fis, clone NT2RP2004321 |
| NM_002189 | -0.537867 | IL15RA | interleukin 15 receptor, alpha |
| NM_012110 | -0.537562 | CHIC2 | cystein-rich hydrophobic domain 2 |
| AB040881 | -0.537473 | KIAA1448 | KIAA1448 protein |
| NM_016577 | -0.537430 | RAB6B | RAB6B, member RAS oncogene family |
| 35 NM_001745 | 0.536940 | CAMLG | calcium modulating ligand |

| | Identifier | Correlation | Name | Description |
|----|----------------|-------------|----------|--|
| | NM_005742 | -0.536738 | P5 | protein disulfide isomerase-related protein |
| | AB011132 | 0.536345 | KIAA0560 | KIAA0560 gene product |
| 5 | Contig54898_RC | 0.536094 | PNN | pinin, desmosome associated protein |
| | Contig45049_RC | -0.536043 | FUT4 | fucosyltransferase 4 (alpha (1,3) fucosyltransferase, myeloid-specific) |
| | NM_006864 | -0.535924 | LILRB3 | leukocyte immunoglobulin-like receptor, subfamily B (with TM and ITIM domains), member 3 |
| 10 | Contig53242_RC | -0.535909 | | Homo sapiens cDNA FLJ11436 fis, clone HEMBA1001213 |
| | NM_005544 | 0.535712 | IRS1 | insulin receptor substrate 1 |
| | Contig47456_RC | 0.535493 | CACNA1D | calcium channel, voltage-dependent, L type, alpha 1D subunit |
| | Contig42751_RC | -0.535469 | | ESTs |
| 15 | Contig29126_RC | -0.535186 | | ESTs |
| | NM_012391 | 0.535067 | PDEF | prostate epithelium-specific Ets transcription factor |
| | NM_012429 | 0.534974 | SEC14L2 | SEC14 (S. cerevisiae)-like 2 |
| | NM_018171 | 0.534898 | FLJ10659 | hypothetical protein FLJ10659 |
| | Contig53047_RC | -0.534773 | TTYH1 | tweety (Drosophila) homolog 1 |
| 20 | Contig54968_RC | 0.534754 | | Homo sapiens cDNA FLJ13558 fis, clone PLACE1007743 |
| | Contig2099_RC | -0.534694 | KIAA1691 | KIAA1691 protein |
| | NM_005264 | 0.534057 | GFRA1 | GDNF family receptor alpha 1 |
| | NM_014036 | -0.533638 | SBB142 | BCM-like membrane protein precursor |
| 25 | NM_018101 | -0.533473 | FLJ10468 | hypothetical protein FLJ10468 |
| | Contig56765_RC | 0.533442 | | ESTs, Moderately similar to K02E10.2 [C.elegans] |
| | AB006746 | -0.533400 | PLSCR1 | phospholipid scramblase 1 |
| | NM_001089 | 0.533350 | ABCA3 | ATP-binding cassette, sub-family A (ABC1), member 3 |
| 30 | NM_018188 | -0.533132 | FLJ10709 | hypothetical protein FLJ10709 |
| | X94232 | -0.532925 | MAPRE2 | microtubule-associated protein, RP/EB family, member 2 |
| | AF234532 | -0.532910 | MYO10 | myosin X |
| | Contig292_RC | 0.532853 | FLJ22386 | hypothetical protein FLJ22386 |
| 35 | NM_000101 | -0.532767 | CYBA | cytochrome b-245, alpha polypeptide |
| | Contig47814_RC | -0.532656 | HHGP | HHGP protein |

| Identifier | Correlation | Name | Description |
|-------------------|-------------|----------|--|
| NM_014320 | -0.532430 | SOUL | putative heme-binding protein |
| NM_020347 | 0.531976 | LZTFL1 | leucine zipper transcription factor-like 1 |
| 5 NM_004323 | 0.531936 | BAG1 | BCL2-associated athanogene |
| Contig50850_RC | -0.531914 | | ESTs |
| Contig11648_RC | 0.531704 | | ESTs |
| NM_018131 | -0.531559 | FLJ10540 | hypothetical protein FLJ10540 |
| NM_004688 | -0.531329 | NMI | N-myc (and STAT) interactor |
| NM_014870 | 0.531101 | KIAA0478 | KIAA0478 gene product |
| 10 Contig31424_RC | 0.530720 | | ESTs |
| NM_000874 | -0.530545 | IFNAR2 | interferon (alpha, beta and omega) receptor 2 |
| Contig50588_RC | 0.530145 | | ESTs |
| NM_016463 | 0.529998 | HSPC195 | hypothetical protein |
| 15 NM_013324 | 0.529966 | CISH | cytokine inducible SH2-containing protein |
| NM_006705 | 0.529840 | GADD45G | growth arrest and DNA-damage-inducible, gamma |
| Contig38901_RC | -0.529747 | | ESTs |
| NM_004184 | -0.529635 | WARS | tryptophanyl-tRNA synthetase |
| NM_015955 | -0.529538 | LOC51072 | CGI-27 protein |
| 20 AF151810 | 0.529416 | CGI-52 | similar to phosphatidylcholine transfer protein 2 |
| NM_002164 | -0.529117 | INDO | indoleamine-pyrrole 2,3 dioxygenase |
| NM_004267 | -0.528679 | CHST2 | carbohydrate (chondroitin 6/keratan) sulfotransferase 2 |
| 25 Contig32185_RC | -0.528529 | | Homo sapiens cDNA FLJ13997 fis, clone Y79AA1002220 |
| NM_004154 | -0.528343 | P2RY6 | pyrimidinergic receptor P2Y, G-protein coupled, 6 |
| NM_005235 | 0.528294 | ERBB4 | v-erb-a avian erythroblastic leukemia viral oncogene homolog-like 4 |
| 30 Contig40208_RC | -0.528062 | LOC56938 | transcription factor BMAL2 |
| NM_013262 | 0.527297 | MIR | myosin regulatory light chain interacting protein |
| NM_003034 | -0.527148 | SIAT8A | sialyltransferase 8 (alpha-N-acetylneuraminate: alpha-2,8-sialyltransferase, GD3 synthase) A |

35

| Identifier | Correlation | Name | Description |
|-------------------|-------------|----------|---|
| NM_004556 | -0.527146 | NFKBIE | nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, epsilon |
| 5 NM_002046 | -0.527051 | GAPD | glyceraldehyde-3-phosphate dehydrogenase |
| NM_001905 | -0.526986 | CTPS | CTP synthase |
| Contig42402_RC | 0.526852 | | ESTs |
| NM_014272 | -0.526283 | ADAMTS7 | a disintegrin-like and metalloprotease (repolysin type) with thrombospondin type 1 motif, 7 |
| 10 AF076612 | 0.526205 | CHRD | chordin |
| Contig57725_RC | -0.526122 | | Homo sapiens mRNA for HMG-box transcription factor TCF-3, complete cds |
| Contig42041_RC | -0.525877 | | ESTs |
| 15 Contig44656_RC | -0.525868 | | ESTs, Highly similar to S02392 alpha-2-macroglobulin receptor precursor [H.sapiens] |
| NM_018004 | -0.525610 | FLJ10134 | hypothetical protein FLJ10134 |
| Contig56434_RC | 0.525510 | | Homo sapiens cDNA FLJ13603 fis, clone PLACE1010270 |
| D25328 | -0.525504 | PFKP | phosphofructokinase, platelet |
| 20 Contig55950_RC | -0.525358 | FLJ22329 | hypothetical protein FLJ22329 |
| NM_002648 | -0.525211 | PIM1 | pim-1 oncogene |
| AL157505 | 0.525186 | | Homo sapiens mRNA; cDNA DKFZp586P1124 (from clone DKFZp586P1124) |
| AF061034 | -0.525185 | FIP2 | Homo sapiens FIP2 alternatively translated mRNA, complete cds. |
| 25 NM_014721 | -0.525102 | KIAA0680 | KIAA0680 gene product |
| NM_001634 | -0.525030 | AMD1 | S-adenosylmethionine decarboxylase 1 |
| NM_006304 | -0.524911 | DSS1 | Deleted in split-hand/split-foot 1 region |
| 30 Contig37778_RC | 0.524667 | | ESTs, Highly similar to HLHUSB MHC class II histocompatibility antigen HLA-DP alpha-1 chain precursor [H.sapiens] |
| NM_003099 | 0.524339 | SNX1 | sorting nexin 1 |
| AL079298 | 0.523774 | MCCC2 | methycrotonoyl-Coenzyme A carboxylase 2 (beta) |
| 35 NM_019013 | -0.523663 | FLJ10156 | hypothetical protein |

| Identifier | Correlation | Name | Description |
|----------------|-------------|----------|--|
| NM_000397 | -0.523293 | CYBB | cytochrome b-245, beta polypeptide (chronic granulomatous disease) |
| NM_014811 | 0.523132 | KIAA0649 | KIAA0649 gene product |
| Contig20600_RC | 0.523072 | | ESTs |
| NM_005190 | -0.522710 | CCNC | cyclin C |
| AL161960 | -0.522574 | FLJ21324 | hypothetical protein FLJ21324 |
| AL117502 | 0.522280 | | Homo sapiens mRNA; cDNA DKFZp434D0935 (from clone DKFZp434D0935) |
| AF131753 | -0.522245 | | Homo sapiens clone 24859 mRNA sequence |
| NM_000320 | 0.521974 | QDPR | quinoid dihydropteridine reductase |
| NM_002115 | -0.521870 | HK3 | hexokinase 3 (white cell) |
| NM_006460 | 0.521696 | HIS1 | HMBA-inducible |
| NM_018683 | -0.521679 | ZNF313 | zinc finger protein 313 |
| NM_004305 | -0.521539 | BIN1 | bridging integrator 1 |
| NM_006770 | -0.521538 | MARCO | macrophage receptor with collagenous structure |
| NM_001166 | -0.521530 | BIRC2 | baculoviral IAP repeat-containing 2 |
| D42047 | 0.521522 | KIAA0089 | KIAA0089 protein |
| NM_016235 | -0.521298 | GPRC5B | G protein-coupled receptor, family C, group 5, member B |
| NM_004504 | -0.521189 | HRB | HIV-1 Rev binding protein |
| NM_002727 | -0.521146 | PRG1 | proteoglycan 1, secretory granule |
| AB029031 | -0.520761 | KIAA1108 | KIAA1108 protein |
| NM_005556 | -0.520692 | KRT7 | keratin 7 |
| NM_018031 | 0.520600 | WDR6 | WD repeat domain 6 |
| AL117523 | -0.520579 | KIAA1053 | KIAA1053 protein |
| NM_004515 | -0.520363 | ILF2 | interleukin enhancer binding factor 2, 45kD |
| NM_004708 | -0.519935 | PDCD5 | programmed cell death 5 |
| NM_005935 | 0.519765 | MLLT2 | myeloid/lymphoid or mixed-lineage leukemia (trithorax (Drosophila) homolog); translocated to, 2 |
| Contig49289_RC | -0.519546 | | Homo sapiens mRNA; cDNA DKFZp586J1119 (from clone DKFZp586J1119); complete cds |
| NM_000211 | -0.519342 | ITGB2 | integrin, beta 2 (antigen CD18 (p95), lymphocyte function-associated antigen 1; macrophage antigen 1 (mac-1) beta subunit) |

| Identifier | Correlation | Name | Description |
|----------------|-------------|---------------|--|
| AL079276 | 0.519207 | LOC58495 | putative zinc finger protein from EUROIMAGE 566589 |
| Contig57825_RC | 0.519041 | | ESTs |
| 5 NM_002466 | -0.518911 | MYBL2 | v-myb avian myeloblastosis viral oncogene homolog-like 2 |
| NM_016072 | -0.518802 | LOC51026 | CGI-141 protein |
| AB007950 | -0.518699 | KIAA0481 | KIAA0481 gene product |
| NM_001550 | -0.518549 | IFRD1 | interferon-related developmental regulator 1 |
| 10 AF155120 | -0.518221 | UBE2V1 | ubiquitin-conjugating enzyme E2 variant 1 |
| Contig49849_RC | 0.517983 | | ESTs, Weakly similar to AF188706 1 g20 protein [H.sapiens] |
| NM_016625 | -0.517936 | LOC51319 | hypothetical protein |
| NM_004049 | -0.517862 | BCL2A1 | BCL2-related protein A1 |
| Contig50719_RC | 0.517740 | | ESTs |
| 15 D80010 | -0.517620 | LPIN1 | lipin 1 |
| NM_000299 | -0.517405 | PKP1 | plakophilin 1 (ectodermal dysplasia/skin fragility syndrome) |
| AL049365 | 0.517080 | FTL | ferritin, light polypeptide |
| Contig65227 | 0.517003 | | ESTs |
| 20 NM_004865 | -0.516808 | TBPL1 | TBP-like 1 |
| Contig54813_RC | 0.516246 | FLJ13962 | hypothetical protein FLJ13962 |
| NM_003494 | -0.516221 | DYSF | dysferlin, limb girdle muscular dystrophy 2B (autosomal recessive) |
| NM_004431 | -0.516212 | EPHA2 | EphA2 |
| AL117600 | -0.516067 | DKFZP564J0863 | DKFZP564J0863 protein |
| 25 AL080209 | -0.516037 | DKFZP586F2423 | hypothetical protein DKFZp586F2423 |
| NM_000135 | -0.515613 | FANCA | Fanconi anemia, complementation group A |
| NM_000050 | -0.515494 | ASS | argininosuccinate synthetase |
| 30 NM_001830 | -0.515439 | CLCN4 | chloride channel 4 |
| NM_018234 | -0.515365 | FLJ10829 | hypothetical protein FLJ10829 |
| Contig53307_RC | 0.515328 | | ESTs, Highly similar to KIAA1437 protein [H.sapiens] |
| AL117617 | -0.515141 | | Homo sapiens mRNA; cDNA DKFZp564H0764 (from clone DKFZp564H0764) |
| 35 NM_002906 | -0.515098 | RDX | radixin |

| Identifier | Correlation | Name | Description |
|-------------------|-------------|----------|--|
| NM_003360 | -0.514427 | UGT8 | UDP glycosyltransferase 8 (UDP-galactose ceramide galactosyltransferase) |
| 5 NM_018478 | 0.514332 | HSMNP1 | uncharacterized hypothalamus protein HSMNP1 |
| M90657 | -0.513908 | TM4SF1 | transmembrane 4 superfamily member 1 |
| NM_014967 | 0.513793 | KIAA1018 | KIAA1018 protein |
| Contig1462_RC | 0.513604 | C11ORF15 | chromosome 11 open reading frame 15 |
| 10 Contig37287_RC | -0.513324 | | ESTs |
| NM_000355 | -0.513225 | TCN2 | transcobalamin II; macrocytic anemia |
| AB037756 | 0.512914 | KIAA1335 | hypothetical protein KIAA1335 |
| Contig842_RC | -0.512880 | | ESTs |
| NM_018186 | -0.512878 | FLJ10706 | hypothetical protein FLJ10706 |
| 15 NM_014668 | 0.512746 | KIAA0575 | KIAA0575 gene product |
| NM_003226 | 0.512611 | TFF3 | trefoil factor 3 (intestinal) |
| Contig56457_RC | -0.512548 | TMEFF1 | transmembrane protein with EGF-like and two follistatin-like domains 1 |
| 20 AL050367 | -0.511999 | | Homo sapiens mRNA; cDNA DKFZp564A026 (from clone DKFZp564A026) |
| NM_014791 | -0.511963 | KIAA0175 | KIAA0175 gene product |
| Contig36312_RC | 0.511794 | | ESTs |
| NM_004811 | -0.511447 | LPXN | leupaxin |
| Contig67182_RC | -0.511416 | | ESTs, Highly similar to epithelial V-like antigen precursor [H.sapiens] |
| 25 Contig52723_RC | -0.511134 | | ESTs |
| Contig17105_RC | -0.511072 | | Homo sapiens mRNA for putative cytoplasmatic protein (ORF1-FL21) |
| NM_014449 | 0.511023 | A | protein "A" |
| Contig52957_RC | 0.510815 | | ESTs |
| Contig49388_RC | 0.510582 | FLJ13322 | hypothetical protein FLJ13322 |
| 30 NM_017786 | 0.510557 | FLJ20366 | hypothetical protein FLJ20366 |
| AL157476 | 0.510478 | | Homo sapiens mRNA; cDNA DKFZp761C082 (from clone DKFZp761C082) |
| 35 NM_001919 | 0.510242 | DCI | dodecenoyl-Coenzyme A delta isomerase (3,2 trans-enoyl-Coenzyme A isomerase) |

| Identifier | Correlation | Name | Description |
|----------------|-------------|----------|--|
| NM_000268 | -0.510165 | NF2 | neurofibromin 2 (bilateral acoustic neuroma) |
| NM_016210 | 0.510018 | LOC51161 | g20 protein |
| Contig45816_RC | -0.509977 | | ESTs |
| NM_003953 | -0.509969 | MPZL1 | myelin protein zero-like 1 |
| NM_000057 | -0.509669 | BLM | Bloom syndrome |
| NM_014452 | -0.509473 | DR6 | death receptor 6 |
| Contig45156_RC | 0.509284 | | ESTs, Moderately similar to motor domain of KIF12 [M.musculus] |
| NM_006943 | 0.509149 | SOX22 | SRY (sex determining region Y)-box 22 |
| NM_000594 | -0.509012 | TNF | tumor necrosis factor (TNF superfamily, member 2) |
| AL137316 | -0.508353 | KIAA1609 | KIAA1609 protein |
| NM_000557 | -0.508325 | GDF5 | growth differentiation factor 5 (cartilage-derived morphogenetic protein-1) |
| NM_018685 | -0.508307 | ANLN | anillin (Drosophila Scraps homolog), actin binding protein |
| Contig53401_RC | 0.508189 | | ESTs |
| NM_014364 | -0.508170 | GAPDS | glyceraldehyde-3-phosphate dehydrogenase, testis-specific |
| Contig50297_RC | 0.508137 | | ESTs, Moderately similar to ALU8_HUMAN ALU SUBFAMILY SX SEQUENCE CONTAMINATION WARNING ENTRY [H.sapiens] |
| Contig51800 | 0.507891 | | ESTs, Weakly similar to ALU6_HUMAN ALU SUBFAMILY SP SEQUENCE CONTAMINATION WARNING ENTRY [H.sapiens] |
| Contig49098_RC | -0.507716 | MGC4090 | hypothetical protein MGC4090 |
| NM_002985 | -0.507554 | SCYA5 | small inducible cytokine A5 (RANTES) |
| AB007899 | 0.507439 | KIAA0439 | KIAA0439 protein; homolog of yeast ubiquitin-protein ligase Rsp5 |
| AL110139 | 0.507145 | | Homo sapiens mRNA; cDNA DKFZp564O1763 (from clone DKFZp564O1763) |
| Contig51117_RC | 0.507001 | | ESTs |
| NM_017660 | -0.506768 | FLJ20085 | hypothetical protein FLJ20085 |
| NM_018000 | 0.506686 | FLJ10116 | hypothetical protein FLJ10116 |
| NM_005555 | -0.506516 | KRT6B | keratin 6B |

| Identifier | Correlation | Name | Description |
|----------------|-------------|----------|---|
| NM_005582 | -0.506462 | LY64 | lymphocyte antigen 64 (mouse) homolog, radioprotective, 105kD |
| Contig47405_RC | 0.506202 | | ESTs |
| 5 NM_014808 | 0.506173 | KIAA0793 | KIAA0793 gene product |
| NM_004938 | -0.506121 | DAPK1 | death-associated protein kinase 1 |
| NM_020659 | -0.505793 | TTYH1 | tweety (Drosophila) homolog 1 |
| NM_006227 | -0.505604 | PLTP | phospholipid transfer protein |
| NM_014268 | -0.505412 | MAPRE2 | microtubule-associated protein, RP/EB family, member 2 |
| 10 NM_004711 | 0.504849 | SYNGR1 | synaptogyrin 1 |
| NM_004418 | -0.504497 | DUSP2 | dual specificity phosphatase 2 |
| NM_003508 | -0.504475 | FZD9 | frizzled (Drosophila) homolog 9 |

15

20

25

30

35

Table 3. 430 gene markers that distinguish *BRCA1*-related tumor samples from sporadic tumor samples

| GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|-----------------------------|---------------|-----------------------------|----------------|
| AB002301 | SEQ ID NO 4 | NM_012391 | SEQ ID NO 1406 |
| AB004857 | SEQ ID NO 8 | NM_012428 | SEQ ID NO 1412 |
| AB007458 | SEQ ID NO 12 | NM_013233 | SEQ ID NO 1418 |
| AB014534 | SEQ ID NO 29 | NM_013253 | SEQ ID NO 1422 |
| AB018305 | SEQ ID NO 34 | NM_013262 | SEQ ID NO 1425 |
| AB020677 | SEQ ID NO 36 | NM_013372 | SEQ ID NO 1434 |
| AB020689 | SEQ ID NO 37 | NM_013378 | SEQ ID NO 1435 |
| AB023151 | SEQ ID NO 41 | NM_014096 | SEQ ID NO 1450 |
| AB023163 | SEQ ID NO 43 | NM_014242 | SEQ ID NO 1464 |
| AB028986 | SEQ ID NO 48 | NM_014314 | SEQ ID NO 1472 |
| AB029025 | SEQ ID NO 50 | NM_014398 | SEQ ID NO 1486 |
| AB032966 | SEQ ID NO 53 | NM_014402 | SEQ ID NO 1488 |
| AB032988 | SEQ ID NO 57 | NM_014476 | SEQ ID NO 1496 |
| AB033049 | SEQ ID NO 63 | NM_014521 | SEQ ID NO 1499 |
| AB033055 | SEQ ID NO 66 | NM_014585 | SEQ ID NO 1504 |
| AB037742 | SEQ ID NO 73 | NM_014597 | SEQ ID NO 1506 |
| AB041269 | SEQ ID NO 96 | NM_014642 | SEQ ID NO 1510 |
| AF000974 | SEQ ID NO 97 | NM_014679 | SEQ ID NO 1517 |
| AF042838 | SEQ ID NO 111 | NM_014680 | SEQ ID NO 1518 |
| AF052155 | SEQ ID NO 119 | NM_014700 | SEQ ID NO 1520 |
| AF055084 | SEQ ID NO 125 | NM_014723 | SEQ ID NO 1523 |
| AF063725 | SEQ ID NO 129 | NM_014770 | SEQ ID NO 1530 |
| AF070536 | SEQ ID NO 133 | NM_014785 | SEQ ID NO 1534 |
| AF070617 | SEQ ID NO 135 | NM_014817 | SEQ ID NO 1539 |
| AF073299 | SEQ ID NO 136 | NM_014840 | SEQ ID NO 1541 |
| AF079529 | SEQ ID NO 140 | NM_014878 | SEQ ID NO 1546 |
| AF090353 | SEQ ID NO 141 | NM_015493 | SEQ ID NO 1564 |
| AF116238 | SEQ ID NO 155 | NM_015523 | SEQ ID NO 1568 |
| AF151810 | SEQ ID NO 171 | NM_015544 | SEQ ID NO 1570 |
| AF220492 | SEQ ID NO 185 | NM_015623 | SEQ ID NO 1572 |
| AJ224741 | SEQ ID NO 196 | NM_015640 | SEQ ID NO 1573 |
| AJ250475 | SEQ ID NO 201 | NM_015721 | SEQ ID NO 1576 |
| AJ270996 | SEQ ID NO 202 | NM_015881 | SEQ ID NO 1577 |
| AJ272057 | SEQ ID NO 203 | NM_015937 | SEQ ID NO 1582 |
| AK000174 | SEQ ID NO 211 | NM_015964 | SEQ ID NO 1586 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|---------------|-----------------------------|----------------|
| | AK000617 | SEQ ID NO 215 | NM_015984 | SEQ ID NO 1587 |
| | AK000959 | SEQ ID NO 222 | NM_016000 | SEQ ID NO 1591 |
| 5 | AK001438 | SEQ ID NO 229 | NM_016018 | SEQ ID NO 1593 |
| | AK001838 | SEQ ID NO 233 | NM_016066 | SEQ ID NO 1601 |
| | AK002107 | SEQ ID NO 238 | NM_016073 | SEQ ID NO 1603 |
| | AK002197 | SEQ ID NO 239 | NM_016081 | SEQ ID NO 1604 |
| | AL035297 | SEQ ID NO 241 | NM_016140 | SEQ ID NO 1611 |
| | AL049346 | SEQ ID NO 243 | NM_016223 | SEQ ID NO 1622 |
| 10 | AL049370 | SEQ ID NO 245 | NM_016267 | SEQ ID NO 1629 |
| | AL049667 | SEQ ID NO 249 | NM_016307 | SEQ ID NO 1633 |
| | AL080222 | SEQ ID NO 276 | NM_016364 | SEQ ID NO 1639 |
| | AL096737 | SEQ ID NO 279 | NM_016373 | SEQ ID NO 1640 |
| | AL110163 | SEQ ID NO 282 | NM_016459 | SEQ ID NO 1646 |
| 15 | AL133057 | SEQ ID NO 300 | NM_016471 | SEQ ID NO 1648 |
| | AL133096 | SEQ ID NO 302 | NM_016548 | SEQ ID NO 1654 |
| | AL133572 | SEQ ID NO 305 | NM_016620 | SEQ ID NO 1662 |
| | AL133619 | SEQ ID NO 307 | NM_016820 | SEQ ID NO 1674 |
| | AL133623 | SEQ ID NO 309 | NM_017423 | SEQ ID NO 1678 |
| | AL137347 | SEQ ID NO 320 | NM_017709 | SEQ ID NO 1698 |
| 20 | AL137381 | SEQ ID NO 322 | NM_017732 | SEQ ID NO 1700 |
| | AL137461 | SEQ ID NO 325 | NM_017734 | SEQ ID NO 1702 |
| | AL137540 | SEQ ID NO 328 | NM_017750 | SEQ ID NO 1704 |
| | AL137555 | SEQ ID NO 329 | NM_017763 | SEQ ID NO 1706 |
| | AL137638 | SEQ ID NO 332 | NM_017782 | SEQ ID NO 1710 |
| 25 | AL137639 | SEQ ID NO 333 | NM_017816 | SEQ ID NO 1714 |
| | AL137663 | SEQ ID NO 334 | NM_018043 | SEQ ID NO 1730 |
| | AL137761 | SEQ ID NO 339 | NM_018072 | SEQ ID NO 1734 |
| | AL157431 | SEQ ID NO 340 | NM_018093 | SEQ ID NO 1738 |
| | AL161960 | SEQ ID NO 351 | NM_018103 | SEQ ID NO 1742 |
| | AL355708 | SEQ ID NO 353 | NM_018171 | SEQ ID NO 1751 |
| 30 | AL359053 | SEQ ID NO 354 | NM_018187 | SEQ ID NO 1755 |
| | D26488 | SEQ ID NO 359 | NM_018188 | SEQ ID NO 1756 |
| | D38521 | SEQ ID NO 361 | NM_018222 | SEQ ID NO 1761 |
| | D50914 | SEQ ID NO 367 | NM_018228 | SEQ ID NO 1762 |
| | D80001 | SEQ ID NO 369 | NM_018373 | SEQ ID NO 1777 |
| 35 | G26403 | SEQ ID NO 380 | NM_018390 | SEQ ID NO 1781 |
| | K02276 | SEQ ID NO 383 | NM_018422 | SEQ ID NO 1784 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|---------------|-----------------------------|----------------|
| | M21551 | SEQ ID NO 394 | NM_018509 | SEQ ID NO 1792 |
| | M27749 | SEQ ID NO 397 | NM_018584 | SEQ ID NO 1796 |
| 5 | M28170 | SEQ ID NO 398 | NM_018653 | SEQ ID NO 1797 |
| | M73547 | SEQ ID NO 409 | NM_018660 | SEQ ID NO 1798 |
| | M80899 | SEQ ID NO 411 | NM_018683 | SEQ ID NO 1799 |
| | NM_000067 | SEQ ID NO 423 | NM_019049 | SEQ ID NO 1814 |
| | NM_000087 | SEQ ID NO 427 | NM_019063 | SEQ ID NO 1815 |
| | NM_000090 | SEQ ID NO 428 | NM_020150 | SEQ ID NO 1823 |
| 10 | NM_000165 | SEQ ID NO 444 | NM_020987 | SEQ ID NO 1848 |
| | NM_000168 | SEQ ID NO 445 | NM_021095 | SEQ ID NO 1855 |
| | NM_000196 | SEQ ID NO 449 | NM_021242 | SEQ ID NO 1867 |
| | NM_000269 | SEQ ID NO 457 | U41387 | SEQ ID NO 1877 |
| | NM_000310 | SEQ ID NO 466 | U45975 | SEQ ID NO 1878 |
| 15 | NM_000396 | SEQ ID NO 479 | U58033 | SEQ ID NO 1881 |
| | NM_000397 | SEQ ID NO 480 | U67784 | SEQ ID NO 1884 |
| | NM_000597 | SEQ ID NO 502 | U68385 | SEQ ID NO 1885 |
| | NM_000636 | SEQ ID NO 509 | U80736 | SEQ ID NO 1890 |
| | NM_000888 | SEQ ID NO 535 | X00437 | SEQ ID NO 1899 |
| | NM_000903 | SEQ ID NO 536 | X07203 | SEQ ID NO 1904 |
| 20 | NM_000930 | SEQ ID NO 540 | X16302 | SEQ ID NO 1907 |
| | NM_000931 | SEQ ID NO 541 | X51630 | SEQ ID NO 1908 |
| | NM_000969 | SEQ ID NO 547 | X57809 | SEQ ID NO 1912 |
| | NM_000984 | SEQ ID NO 548 | X57819 | SEQ ID NO 1913 |
| | NM_001026 | SEQ ID NO 552 | X58529 | SEQ ID NO 1914 |
| 25 | NM_001054 | SEQ ID NO 554 | X66087 | SEQ ID NO 1916 |
| | NM_001179 | SEQ ID NO 567 | X69150 | SEQ ID NO 1917 |
| | NM_001184 | SEQ ID NO 568 | X72475 | SEQ ID NO 1918 |
| | NM_001204 | SEQ ID NO 571 | X74794 | SEQ ID NO 1920 |
| | NM_001206 | SEQ ID NO 572 | X75315 | SEQ ID NO 1921 |
| | NM_001218 | SEQ ID NO 575 | X84340 | SEQ ID NO 1925 |
| 30 | NM_001275 | SEQ ID NO 586 | X98260 | SEQ ID NO 1928 |
| | NM_001394 | SEQ ID NO 602 | Y07512 | SEQ ID NO 1931 |
| | NM_001424 | SEQ ID NO 605 | Y14737 | SEQ ID NO 1932 |
| | NM_001448 | SEQ ID NO 610 | Z34893 | SEQ ID NO 1934 |
| | NM_001504 | SEQ ID NO 620 | Contig237_RC | SEQ ID NO 1940 |
| 35 | NM_001553 | SEQ ID NO 630 | Contig292_RC | SEQ ID NO 1942 |
| | NM_001674 | SEQ ID NO 646 | Contig372_RC | SEQ ID NO 1943 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|---------------|-----------------------------|----------------|
| | NM_001675 | SEQ ID NO 647 | Contig756_RC | SEQ ID NO 1955 |
| | NM_001725 | SEQ ID NO 652 | Contig842_RC | SEQ ID NO 1958 |
| 5 | NM_001740 | SEQ ID NO 656 | Contig1632_RC | SEQ ID NO 1977 |
| | NM_001756 | SEQ ID NO 659 | Contig1826_RC | SEQ ID NO 1980 |
| | NM_001770 | SEQ ID NO 664 | Contig2237_RC | SEQ ID NO 1988 |
| | NM_001797 | SEQ ID NO 670 | Contig2915_RC | SEQ ID NO 2003 |
| | NM_001845 | SEQ ID NO 680 | Contig3164_RC | SEQ ID NO 2007 |
| | NM_001873 | SEQ ID NO 684 | Contig3252_RC | SEQ ID NO 2008 |
| 10 | NM_001888 | SEQ ID NO 687 | Contig3940_RC | SEQ ID NO 2018 |
| | NM_001892 | SEQ ID NO 688 | Contig9259_RC | SEQ ID NO 2039 |
| | NM_001919 | SEQ ID NO 694 | Contig10268_RC | SEQ ID NO 2041 |
| | NM_001946 | SEQ ID NO 698 | Contig10437_RC | SEQ ID NO 2043 |
| | NM_001953 | SEQ ID NO 699 | Contig10973_RC | SEQ ID NO 2044 |
| 15 | NM_001960 | SEQ ID NO 704 | Contig14390_RC | SEQ ID NO 2054 |
| | NM_001985 | SEQ ID NO 709 | Contig16453_RC | SEQ ID NO 2060 |
| | NM_002023 | SEQ ID NO 712 | Contig16759_RC | SEQ ID NO 2061 |
| | NM_002051 | SEQ ID NO 716 | Contig19551 | SEQ ID NO 2070 |
| | NM_002053 | SEQ ID NO 717 | Contig24541_RC | SEQ ID NO 2088 |
| | NM_002164 | SEQ ID NO 734 | Contig25362_RC | SEQ ID NO 2093 |
| 20 | NM_002200 | SEQ ID NO 739 | Contig25617_RC | SEQ ID NO 2094 |
| | NM_002201 | SEQ ID NO 740 | Contig25722_RC | SEQ ID NO 2096 |
| | NM_002213 | SEQ ID NO 741 | Contig26022_RC | SEQ ID NO 2099 |
| | NM_002250 | SEQ ID NO 747 | Contig27915_RC | SEQ ID NO 2114 |
| | NM_002512 | SEQ ID NO 780 | Contig28081_RC | SEQ ID NO 2116 |
| 25 | NM_002542 | SEQ ID NO 784 | Contig28179_RC | SEQ ID NO 2118 |
| | NM_002561 | SEQ ID NO 786 | Contig28550_RC | SEQ ID NO 2119 |
| | NM_002615 | SEQ ID NO 793 | Contig29639_RC | SEQ ID NO 2127 |
| | NM_002686 | SEQ ID NO 803 | Contig29647_RC | SEQ ID NO 2128 |
| | NM_002709 | SEQ ID NO 806 | Contig30092_RC | SEQ ID NO 2130 |
| | NM_002742 | SEQ ID NO 812 | Contig30209_RC | SEQ ID NO 2132 |
| 30 | NM_002775 | SEQ ID NO 815 | Contig32185_RC | SEQ ID NO 2156 |
| | NM_002975 | SEQ ID NO 848 | Contig32798_RC | SEQ ID NO 2161 |
| | NM_002982 | SEQ ID NO 849 | Contig33230_RC | SEQ ID NO 2163 |
| | NM_003104 | SEQ ID NO 870 | Contig33394_RC | SEQ ID NO 2165 |
| | NM_003118 | SEQ ID NO 872 | Contig36323_RC | SEQ ID NO 2197 |
| 35 | NM_003144 | SEQ ID NO 876 | Contig36761_RC | SEQ ID NO 2201 |
| | NM_003165 | SEQ ID NO 882 | Contig37141_RC | SEQ ID NO 2209 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|----------------|-----------------------------|----------------|
| | NM_003197 | SEQ ID NO 885 | Contig37778_RC | SEQ ID NO 2218 |
| | NM_003202 | SEQ ID NO 886 | Contig38285_RC | SEQ ID NO 2222 |
| 5 | NM_003217 | SEQ ID NO 888 | Contig38520_RC | SEQ ID NO 2225 |
| | NM_003283 | SEQ ID NO 898 | Contig38901_RC | SEQ ID NO 2232 |
| | NM_003462 | SEQ ID NO 911 | Contig39826_RC | SEQ ID NO 2241 |
| | NM_003500 | SEQ ID NO 918 | Contig40212_RC | SEQ ID NO 2251 |
| | NM_003561 | SEQ ID NO 925 | Contig40712_RC | SEQ ID NO 2257 |
| | NM_003607 | SEQ ID NO 930 | Contig41402_RC | SEQ ID NO 2265 |
| 10 | NM_003633 | SEQ ID NO 933 | Contig41635_RC | SEQ ID NO 2272 |
| | NM_003641 | SEQ ID NO 934 | Contig42006_RC | SEQ ID NO 2280 |
| | NM_003683 | SEQ ID NO 943 | Contig42220_RC | SEQ ID NO 2286 |
| | NM_003729 | SEQ ID NO 949 | Contig42306_RC | SEQ ID NO 2287 |
| | NM_003793 | SEQ ID NO 954 | Contig43918_RC | SEQ ID NO 2312 |
| 15 | NM_003829 | SEQ ID NO 958 | Contig44195_RC | SEQ ID NO 2316 |
| | NM_003866 | SEQ ID NO 961 | Contig44265_RC | SEQ ID NO 2318 |
| | NM_003904 | SEQ ID NO 967 | Contig44278_RC | SEQ ID NO 2319 |
| | NM_003953 | SEQ ID NO 974 | Contig44757_RC | SEQ ID NO 2329 |
| | NM_004024 | SEQ ID NO 982 | Contig45588_RC | SEQ ID NO 2349 |
| | NM_004053 | SEQ ID NO 986 | Contig46262_RC | SEQ ID NO 2361 |
| 20 | NM_004295 | SEQ ID NO 1014 | Contig46288_RC | SEQ ID NO 2362 |
| | NM_004438 | SEQ ID NO 1038 | Contig46343_RC | SEQ ID NO 2363 |
| | NM_004559 | SEQ ID NO 1057 | Contig46452_RC | SEQ ID NO 2366 |
| | NM_004616 | SEQ ID NO 1065 | Contig46868_RC | SEQ ID NO 2373 |
| | NM_004741 | SEQ ID NO 1080 | Contig46937_RC | SEQ ID NO 2377 |
| 25 | NM_004772 | SEQ ID NO 1084 | Contig48004_RC | SEQ ID NO 2393 |
| | NM_004791 | SEQ ID NO 1086 | Contig48249_RC | SEQ ID NO 2397 |
| | NM_004848 | SEQ ID NO 1094 | Contig48774_RC | SEQ ID NO 2405 |
| | NM_004866 | SEQ ID NO 1097 | Contig48913_RC | SEQ ID NO 2411 |
| | NM_005128 | SEQ ID NO 1121 | Contig48945_RC | SEQ ID NO 2412 |
| | NM_005148 | SEQ ID NO 1124 | Contig48970_RC | SEQ ID NO 2413 |
| 30 | NM_005196 | SEQ ID NO 1127 | Contig49233_RC | SEQ ID NO 2419 |
| | NM_005326 | SEQ ID NO 1140 | Contig49289_RC | SEQ ID NO 2422 |
| | NM_005518 | SEQ ID NO 1161 | Contig49342_RC | SEQ ID NO 2423 |
| | NM_005538 | SEQ ID NO 1163 | Contig49510_RC | SEQ ID NO 2430 |
| | NM_005557 | SEQ ID NO 1170 | Contig49855 | SEQ ID NO 2440 |
| 35 | NM_005718 | SEQ ID NO 1189 | Contig49948_RC | SEQ ID NO 2442 |
| | NM_005804 | SEQ ID NO 1201 | Contig50297_RC | SEQ ID NO 2451 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|----------------|-----------------------------|----------------|
| | NM_005824 | SEQ ID NO 1203 | Contig50669_RC | SEQ ID NO 2458 |
| | NM_005935 | SEQ ID NO 1220 | Contig50673_RC | SEQ ID NO 2459 |
| 5 | NM_006002 | SEQ ID NO 1225 | Contig50838_RC | SEQ ID NO 2465 |
| | NM_006148 | SEQ ID NO 1249 | Contig51068_RC | SEQ ID NO 2471 |
| | NM_006235 | SEQ ID NO 1257 | Contig51929 | SEQ ID NO 2492 |
| | NM_006271 | SEQ ID NO 1261 | Contig51953_RC | SEQ ID NO 2493 |
| | NM_006287 | SEQ ID NO 1264 | Contig52405_RC | SEQ ID NO 2502 |
| | NM_006296 | SEQ ID NO 1267 | Contig52543_RC | SEQ ID NO 2505 |
| 10 | NM_006378 | SEQ ID NO 1275 | Contig52720_RC | SEQ ID NO 2513 |
| | NM_006461 | SEQ ID NO 1287 | Contig53281_RC | SEQ ID NO 2530 |
| | NM_006573 | SEQ ID NO 1300 | Contig53598_RC | SEQ ID NO 2537 |
| | NM_006622 | SEQ ID NO 1302 | Contig53757_RC | SEQ ID NO 2543 |
| | NM_006696 | SEQ ID NO 1308 | Contig53944_RC | SEQ ID NO 2545 |
| 15 | NM_006769 | SEQ ID NO 1316 | Contig54425 | SEQ ID NO 2561 |
| | NM_006787 | SEQ ID NO 1319 | Contig54547_RC | SEQ ID NO 2565 |
| | NM_006875 | SEQ ID NO 1334 | Contig54757_RC | SEQ ID NO 2574 |
| | NM_006885 | SEQ ID NO 1335 | Contig54916_RC | SEQ ID NO 2581 |
| | NM_006918 | SEQ ID NO 1339 | Contig55770_RC | SEQ ID NO 2604 |
| | NM_006923 | SEQ ID NO 1340 | Contig55801_RC | SEQ ID NO 2606 |
| 20 | NM_006941 | SEQ ID NO 1342 | Contig56143_RC | SEQ ID NO 2619 |
| | NM_007070 | SEQ ID NO 1354 | Contig56160_RC | SEQ ID NO 2620 |
| | NM_007088 | SEQ ID NO 1356 | Contig56303_RC | SEQ ID NO 2626 |
| | NM_007146 | SEQ ID NO 1358 | Contig57023_RC | SEQ ID NO 2639 |
| | NM_007173 | SEQ ID NO 1359 | Contig57138_RC | SEQ ID NO 2644 |
| 25 | NM_007246 | SEQ ID NO 1366 | Contig57609_RC | SEQ ID NO 2657 |
| | NM_007358 | SEQ ID NO 1374 | Contig58301_RC | SEQ ID NO 2667 |
| | NM_012135 | SEQ ID NO 1385 | Contig58512_RC | SEQ ID NO 2670 |
| | NM_012151 | SEQ ID NO 1387 | Contig60393 | SEQ ID NO 2674 |
| | NM_012258 | SEQ ID NO 1396 | Contig60509_RC | SEQ ID NO 2675 |
| | NM_012317 | SEQ ID NO 1399 | Contig61254_RC | SEQ ID NO 2677 |
| 30 | NM_012337 | SEQ ID NO 1403 | Contig62306 | SEQ ID NO 2680 |
| | NM_012339 | SEQ ID NO 1404 | Contig64502 | SEQ ID NO 2689 |

Table 4. 100 preferred markers from Table 3 distinguishing *BRCAl*-related tumors from sporadic tumors.

| Identifier | Correlation | Sequence Name | Description |
|-------------------|-------------|---------------|--|
| 5 NM_001892 | -0.651689 | CSNK1A1 | casein kinase 1, alpha 1 |
| NM_018171 | -0.637696 | FLJ10659 | hypothetical protein FLJ10659 |
| Contig40712_RC | -0.612509 | | ESTs |
| NM_001204 | -0.608470 | BMPR2 | bone morphogenetic protein receptor, type II (serine/threonine kinase) |
| 10 NM_005148 | -0.598612 | UNC119 | unc119 (C.elegans) homolog |
| G26403 | 0.585054 | YWHAH | tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, eta polypeptide |
| NM_015640 | 0.583397 | PAI-RBP1 | PAI-1 mRNA-binding protein |
| Contig9259_RC | 0.581362 | | ESTs |
| 15 AB033049 | -0.578750 | KIAA1223 | KIAA1223 protein |
| NM_015523 | 0.576029 | DKFZP566E144 | small fragment nuclease |
| 20 Contig41402_RC | -0.571650 | | Human DNA sequence from clone RP11-16L21 on chromosome 9. Contains the gene for NADP-dependent leukotriene B4 12-hydroxydehydrogenase, the gene for a novel DnaJ domain protein similar to Drosophila, C. elegans and Arabidopsis predicted proteins, the GNG10 gene for guanine nucleotide binding protein 10, a novel gene, ESTs, STSs, GSSs and six CpG islands |
| 25 NM_004791 | -0.564819 | ITGBL1 | integrin, beta-like 1 (with EGF-like repeat domains) |
| NM_007070 | 0.561173 | FAP48 | FKBP-associated protein |
| NM_014597 | 0.555907 | HSU15552 | acidic 82 kDa protein mRNA |
| 30 AF000974 | 0.547194 | TRIP6 | thyroid hormone receptor interactor 6 |
| NM_016073 | -0.547072 | CGI-142 | CGI-142 |
| Contig3940_RC | 0.544073 | YWHAH | tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, eta polypeptide |
| 35 NM_003683 | 0.542219 | D21S2056E | DNA segment on chromosome 21 (unique) 2056 expressed sequence |

| Identifier | Correlation | Sequence Name | Description |
|-------------------|-------------|---------------|---|
| Contig58512_RC | -0.528458 | | Homo sapiens pancreas tumor-related protein (FKSG12) mRNA, complete cds |
| 5 NM_003904 | 0.521223 | ZNF259 | zinc finger protein 259 |
| Contig26022_RC | 0.517351 | | ESTs |
| Contig48970_RC | -0.516953 | KIAA0892 | KIAA0892 protein |
| NM_016307 | -0.515398 | PRX2 | paired related homeobox protein |
| 10 AL137761 | -0.514891 | | Homo sapiens mRNA; cDNA DKFZp586L2424 (from clone DKFZp586L2424) |
| NM_001919 | -0.514799 | DCI | dodecenoyl-Coenzyme A delta isomerase (3,2 trans-enoyl-Coenzyme A isomerase) |
| NM_000196 | -0.514004 | HSD11B2 | hydroxysteroid (11-beta) dehydrogenase 2 |
| 15 NM_002200 | 0.513149 | IRF5 | interferon regulatory factor 5 |
| AL133572 | 0.511340 | | Homo sapiens mRNA; cDNA DKFZp434I0535 (from clone DKFZp434I0535); partial cds |
| NM_019063 | 0.511127 | C2ORF2 | chromosome 2 open reading frame 2 |
| Contig25617_RC | 0.509506 | | ESTs |
| 20 NM_007358 | 0.508145 | M96 | putative DNA binding protein |
| NM_014785 | -0.507114 | KIAA0258 | KIAA0258 gene product |
| NM_006235 | 0.506585 | POU2AF1 | POU domain, class 2, associating factor 1 |
| NM_014680 | -0.505779 | KIAA0100 | KIAA0100 gene product |
| 25 X66087 | 0.500842 | MYBL1 | v-myb avian myeloblastosis viral oncogene homolog-like 1 |
| Y07512 | -0.500686 | PRKG1 | protein kinase, cGMP-dependent, type I |
| NM_006296 | 0.500344 | VRK2 | vaccinia related kinase 2 |
| Contig44278_RC | 0.498260 | DKFZP434K114 | DKFZP434K114 protein |
| 30 Contig56160_RC | -0.497695 | | ESTs |
| NM_002023 | -0.497570 | FMOD | fibromodulin |
| M28170 | 0.497095 | CD19 | CD19 antigen |
| D26488 | 0.496511 | KIAA0007 | KIAA0007 protein |
| 35 X72475 | 0.496125 | | H.sapiens mRNA for rearranged Ig kappa light chain variable region (I.114) |

| | Identifier | Correlation | Sequence Name | Description |
|----|----------------|-------------|---------------|---|
| | K02276 | 0.496068 | MYC | v-myc avian myelocytomatosis viral oncogene homolog |
| 5 | NM_013378 | 0.495648 | VPREB3 | pre-B lymphocyte gene 3 |
| | X58529 | 0.495608 | IGHM | immunoglobulin heavy constant mu |
| | NM_000168 | -0.494260 | GLI3 | GLI-Kruppel family member GLI3 (Greig cephalopolysyndactyly syndrome) |
| | NM_004866 | -0.492967 | SCAMP1 | secretory carrier membrane protein 1 |
| 10 | NM_013253 | -0.491159 | DKK3 | dickkopf (Xenopus laevis) homolog 3 |
| | NM_003729 | 0.488971 | RPC | RNA 3'-terminal phosphate cyclase |
| | NM_006875 | 0.487407 | PIM2 | pim-2 oncogene |
| | NM_018188 | 0.487126 | FLJ10709 | hypothetical protein FLJ10709 |
| | NM_004848 | 0.485408 | ICB-1 | basement membrane-induced gene |
| 15 | NM_001179 | 0.483253 | ART3 | ADP-ribosyltransferase 3 |
| | NM_016548 | -0.482329 | LOC51280 | golgi membrane protein GP73 |
| | NM_007146 | -0.481994 | ZNF161 | zinc finger protein 161 |
| | NM_021242 | -0.481754 | STRAIT11499 | hypothetical protein STRAIT11499 |
| 20 | NM_016223 | 0.481710 | PACSIN3 | protein kinase C and casein kinase substrate in neurons 3 |
| | NM_003197 | -0.481526 | TCEB1L | transcription elongation factor B (SIII), polypeptide 1-like |
| | NM_000067 | -0.481003 | CA2 | carbonic anhydrase II |
| | NM_006885 | -0.479705 | ATBF1 | AT-binding transcription factor 1 |
| | NM_002542 | 0.478282 | OGG1 | 8-oxoguanine DNA glycosylase |
| 25 | AL133619 | -0.476596 | | Homo sapiens mRNA; cDNA DKFZp434E2321 (from clone DKFZp434E2321); partial cds |
| | D80001 | 0.476130 | KIAA0179 | KIAA0179 protein |
| | NM_018660 | -0.475548 | LOC55893 | papillomavirus regulatory factor PRF-1 |
| 30 | AB004857 | 0.473440 | SLC11A2 | solute carrier family 11 (proton-coupled divalent metal ion transporters), member 2 |
| | NM_002250 | 0.472900 | KCNN4 | potassium intermediate/small conductance calcium-activated channel, subfamily N, member 4 |
| 35 | Contig56143_RC | -0.472611 | | ESTs, Weakly similar to A54849 collagen alpha 1(VII) chain precursor [H.sapiens] |

| Identifier | Correlation | Sequence Name | Description |
|-------------------|-------------|---------------|---|
| NM_001960 | 0.471502 | EEF1D | eukaryotic translation elongation factor 1 delta (guanine nucleotide exchange protein) |
| 5 Contig52405_RC | -0.470705 | | ESTs, Weakly similar to ALU8_HUMAN ALU SUBFAMILY SX SEQUENCE CONTAMINATION WARNING ENTRY [H.sapiens] |
| Contig30092_RC | -0.469977 | | Homo sapiens PR-domain zinc finger protein 6 isoform B (PRDM6) mRNA, partial cds; alternatively spliced |
| 10 NM_003462 | -0.468753 | P28 | dynein, axonemal, light intermediate polypeptide |
| Contig60393 | 0.468475 | | ESTs |
| Contig842_RC | 0.468158 | | ESTs |
| 15 NM_002982 | 0.466362 | SCYA2 | small inducible cytokine A2 (monocyte chemotactic protein 1, homologous to mouse Sig-je) |
| Contig14390_RC | 0.464150 | | ESTs |
| NM_001770 | 0.463847 | CD19 | CD19 antigen |
| AK000617 | -0.463158 | | Homo sapiens mRNA; cDNA DKFZp434L235 (from clone DKFZp434L235) |
| 20 AF073299 | -0.463007 | SLC9A2 | solute carrier family 9 (sodium/hydrogen exchanger), isoform 2 |
| NM_019049 | 0.461990 | FLJ20054 | hypothetical protein |
| AL137347 | -0.460778 | DKFZP761M1511 | hypothetical protein |
| 25 NM_000396 | -0.460263 | CTSK | cathepsin K (pseudodysostosis) |
| NM_018373 | -0.459268 | FLJ11271 | hypothetical protein FLJ11271 |
| NM_002709 | 0.458500 | PPP1CB | protein phosphatase 1, catalytic subunit, beta isoform |
| NM_016820 | 0.457516 | OGG1 | 8-oxoguanine DNA glycosylase |
| 30 Contig10268_RC | 0.456933 | | Human DNA sequence from clone RP11-196N14 on chromosome 20 Contains ESTs, STSs, GSSs and CpG islands. Contains three novel genes, part of a gene for a novel protein similar to protein serine/threonine phosphatase 4 regulatory subunit 1 (PP4R1) and a gene for a novel protein with an ankyrin domain |
| 35 | | | |

| Identifier | Correlation | Sequence Name | Description |
|----------------|-------------|---------------|--|
| NM_014521 | -0.456733 | SH3BP4 | SH3-domain binding protein 4 |
| AJ272057 | -0.456548 | STRAIT11499 | hypothetical protein STRAIT11499 |
| 5 NM_015964 | -0.456187 | LOC51673 | brain specific protein |
| Contig16759_RC | -0.456169 | | ESTs |
| NM_015937 | -0.455954 | LOC51604 | CGI-06 protein |
| NM_007246 | -0.455500 | KLHL2 | kelch (Drosophila)-like 2 (Mayven) |
| NM_001985 | -0.453024 | ETFB | electron-transfer-flavoprotein, beta polypeptide |
| 10 NM_000984 | -0.452935 | RPL23A | ribosomal protein L23a |
| Contig51953_RC | -0.451695 | | ESTs |
| NM_015984 | 0.450491 | UCH37 | ubiquitin C-terminal hydrolase UCH37 |
| NM_000903 | -0.450371 | DIA4 | diaphorase (NADH/NADPH) (cytochrome b-5 reductase) |
| 15 NM_001797 | -0.449862 | CDH11 | cadherin 11, type 2, OB-cadherin (osteoblast) |
| NM_014878 | 0.449818 | KIAA0020 | KIAA0020 gene product |
| NM_002742 | -0.449590 | PRKCM | protein kinase C, mu |

20

25

30

35

Table 5. 231 gene markers that distinguish patients with good prognosis from patients with poor prognosis.

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|---------------|-----------------------------|----------------|
| 5 | AA555029_RC | SEQ ID NO 1 | NM_013296 | SEQ ID NO 1427 |
| | AB020689 | SEQ ID NO 37 | NM_013437 | SEQ ID NO 1439 |
| | AB032973 | SEQ ID NO 55 | NM_014078 | SEQ ID NO 1449 |
| | AB033007 | SEQ ID NO 58 | NM_014109 | SEQ ID NO 1451 |
| | AB033043 | SEQ ID NO 62 | NM_014321 | SEQ ID NO 1477 |
| | AB037745 | SEQ ID NO 75 | NM_014363 | SEQ ID NO 1480 |
| 10 | AB037863 | SEQ ID NO 88 | NM_014750 | SEQ ID NO 1527 |
| | AF052159 | SEQ ID NO 120 | NM_014754 | SEQ ID NO 1528 |
| | AF052162 | SEQ ID NO 121 | NM_014791 | SEQ ID NO 1535 |
| | AF055033 | SEQ ID NO 124 | NM_014875 | SEQ ID NO 1545 |
| | AF073519 | SEQ ID NO 137 | NM_014889 | SEQ ID NO 1548 |
| 15 | AF148505 | SEQ ID NO 169 | NM_014968 | SEQ ID NO 1554 |
| | AF155117 | SEQ ID NO 173 | NM_015416 | SEQ ID NO 1559 |
| | AF161553 | SEQ ID NO 177 | NM_015417 | SEQ ID NO 1560 |
| | AF201951 | SEQ ID NO 183 | NM_015434 | SEQ ID NO 1562 |
| | AF257175 | SEQ ID NO 189 | NM_015984 | SEQ ID NO 1587 |
| | AJ224741 | SEQ ID NO 196 | NM_016337 | SEQ ID NO 1636 |
| 20 | AK000745 | SEQ ID NO 219 | NM_016359 | SEQ ID NO 1638 |
| | AL050021 | SEQ ID NO 257 | NM_016448 | SEQ ID NO 1645 |
| | AL050090 | SEQ ID NO 259 | NM_016569 | SEQ ID NO 1655 |
| | AL080059 | SEQ ID NO 270 | NM_016577 | SEQ ID NO 1656 |
| | AL080079 | SEQ ID NO 271 | NM_017779 | SEQ ID NO 1708 |
| 25 | AL080110 | SEQ ID NO 272 | NM_018004 | SEQ ID NO 1725 |
| | AL133603 | SEQ ID NO 306 | NM_018098 | SEQ ID NO 1739 |
| | AL133619 | SEQ ID NO 307 | NM_018104 | SEQ ID NO 1743 |
| | AL137295 | SEQ ID NO 315 | NM_018120 | SEQ ID NO 1745 |
| | AL137502 | SEQ ID NO 326 | NM_018136 | SEQ ID NO 1748 |
| | AL137514 | SEQ ID NO 327 | NM_018265 | SEQ ID NO 1766 |
| 30 | AL137718 | SEQ ID NO 336 | NM_018354 | SEQ ID NO 1774 |
| | AL355708 | SEQ ID NO 353 | NM_018401 | SEQ ID NO 1782 |
| | D25328 | SEQ ID NO 357 | NM_018410 | SEQ ID NO 1783 |
| | L27560 | SEQ ID NO 390 | NM_018454 | SEQ ID NO 1786 |
| | M21551 | SEQ ID NO 394 | NM_018455 | SEQ ID NO 1787 |
| 35 | NM_000017 | SEQ ID NO 416 | NM_019013 | SEQ ID NO 1809 |
| | NM_000096 | SEQ ID NO 430 | NM_020166 | SEQ ID NO 1825 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|---------------|-----------------------------|----------------|
| | NM_000127 | SEQ ID NO 436 | NM_020188 | SEQ ID NO 1830 |
| | NM_000158 | SEQ ID NO 442 | NM_020244 | SEQ ID NO 1835 |
| 5 | NM_000224 | SEQ ID NO 453 | NM_020386 | SEQ ID NO 1838 |
| | NM_000286 | SEQ ID NO 462 | NM_020675 | SEQ ID NO 1842 |
| | NM_000291 | SEQ ID NO 463 | NM_020974 | SEQ ID NO 1844 |
| | NM_000320 | SEQ ID NO 469 | R70506_RC | SEQ ID NO 1868 |
| | NM_000436 | SEQ ID NO 487 | U45975 | SEQ ID NO 1878 |
| | NM_000507 | SEQ ID NO 491 | U58033 | SEQ ID NO 1881 |
| 10 | NM_000599 | SEQ ID NO 503 | U82987 | SEQ ID NO 1891 |
| | NM_000788 | SEQ ID NO 527 | U96131 | SEQ ID NO 1896 |
| | NM_000849 | SEQ ID NO 530 | X05610 | SEQ ID NO 1903 |
| | NM_001007 | SEQ ID NO 550 | X94232 | SEQ ID NO 1927 |
| | NM_001124 | SEQ ID NO 562 | Contig753_RC | SEQ ID NO 1954 |
| 15 | NM_001168 | SEQ ID NO 566 | Contig1778_RC | SEQ ID NO 1979 |
| | NM_001216 | SEQ ID NO 574 | Contig2399_RC | SEQ ID NO 1989 |
| | NM_001280 | SEQ ID NO 588 | Contig2504_RC | SEQ ID NO 1991 |
| | NM_001282 | SEQ ID NO 589 | Contig3902_RC | SEQ ID NO 2017 |
| | NM_001333 | SEQ ID NO 597 | Contig4595 | SEQ ID NO 2022 |
| | NM_001673 | SEQ ID NO 645 | Contig8581_RC | SEQ ID NO 2037 |
| 20 | NM_001809 | SEQ ID NO 673 | Contig13480_RC | SEQ ID NO 2052 |
| | NM_001827 | SEQ ID NO 676 | Contig17359_RC | SEQ ID NO 2068 |
| | NM_001905 | SEQ ID NO 691 | Contig20217_RC | SEQ ID NO 2072 |
| | NM_002019 | SEQ ID NO 711 | Contig21812_RC | SEQ ID NO 2082 |
| | NM_002073 | SEQ ID NO 721 | Contig24252_RC | SEQ ID NO 2087 |
| 25 | NM_002358 | SEQ ID NO 764 | Contig25055_RC | SEQ ID NO 2090 |
| | NM_002570 | SEQ ID NO 787 | Contig25343_RC | SEQ ID NO 2092 |
| | NM_002808 | SEQ ID NO 822 | Contig25991 | SEQ ID NO 2098 |
| | NM_002811 | SEQ ID NO 823 | Contig27312_RC | SEQ ID NO 2108 |
| | NM_002900 | SEQ ID NO 835 | Contig28552_RC | SEQ ID NO 2120 |
| | NM_002916 | SEQ ID NO 838 | Contig32125_RC | SEQ ID NO 2155 |
| 30 | NM_003158 | SEQ ID NO 881 | Contig32185_RC | SEQ ID NO 2156 |
| | NM_003234 | SEQ ID NO 891 | Contig33814_RC | SEQ ID NO 2169 |
| | NM_003239 | SEQ ID NO 893 | Contig34634_RC | SEQ ID NO 2180 |
| | NM_003258 | SEQ ID NO 896 | Contig35251_RC | SEQ ID NO 2185 |
| | NM_003376 | SEQ ID NO 906 | Contig37063_RC | SEQ ID NO 2206 |
| 35 | NM_003600 | SEQ ID NO 929 | Contig37598 | SEQ ID NO 2216 |
| | NM_003607 | SEQ ID NO 930 | Contig38288_RC | SEQ ID NO 2223 |

| | GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|----|-----------------------------|----------------|-----------------------------|----------------|
| | NM_003662 | SEQ ID NO 938 | Contig40128_RC | SEQ ID NO 2248 |
| | NM_003676 | SEQ ID NO 941 | Contig40831_RC | SEQ ID NO 2260 |
| 5 | NM_003748 | SEQ ID NO 951 | Contig41413_RC | SEQ ID NO 2266 |
| | NM_003862 | SEQ ID NO 960 | Contig41887_RC | SEQ ID NO 2276 |
| | NM_003875 | SEQ ID NO 962 | Contig42421_RC | SEQ ID NO 2291 |
| | NM_003878 | SEQ ID NO 963 | Contig43747_RC | SEQ ID NO 2311 |
| | NM_003882 | SEQ ID NO 964 | Contig44064_RC | SEQ ID NO 2315 |
| | NM_003981 | SEQ ID NO 977 | Contig44289_RC | SEQ ID NO 2320 |
| 10 | NM_004052 | SEQ ID NO 985 | Contig44799_RC | SEQ ID NO 2330 |
| | NM_004163 | SEQ ID NO 995 | Contig45347_RC | SEQ ID NO 2344 |
| | NM_004336 | SEQ ID NO 1022 | Contig45816_RC | SEQ ID NO 2351 |
| | NM_004358 | SEQ ID NO 1026 | Contig46218_RC | SEQ ID NO 2358 |
| | NM_004456 | SEQ ID NO 1043 | Contig46223_RC | SEQ ID NO 2359 |
| 15 | NM_004480 | SEQ ID NO 1046 | Contig46653_RC | SEQ ID NO 2369 |
| | NM_004504 | SEQ ID NO 1051 | Contig46802_RC | SEQ ID NO 2372 |
| | NM_004603 | SEQ ID NO 1064 | Contig47405_RC | SEQ ID NO 2384 |
| | NM_004701 | SEQ ID NO 1075 | Contig48328_RC | SEQ ID NO 2400 |
| | NM_004702 | SEQ ID NO 1076 | Contig49670_RC | SEQ ID NO 2434 |
| | NM_004798 | SEQ ID NO 1087 | Contig50106_RC | SEQ ID NO 2445 |
| 20 | NM_004911 | SEQ ID NO 1102 | Contig50410 | SEQ ID NO 2453 |
| | NM_004994 | SEQ ID NO 1108 | Contig50802_RC | SEQ ID NO 2463 |
| | NM_005196 | SEQ ID NO 1127 | Contig51464_RC | SEQ ID NO 2481 |
| | NM_005342 | SEQ ID NO 1143 | Contig51519_RC | SEQ ID NO 2482 |
| | NM_005496 | SEQ ID NO 1157 | Contig51749_RC | SEQ ID NO 2486 |
| 25 | NM_005563 | SEQ ID NO 1173 | Contig51963 | SEQ ID NO 2494 |
| | NM_005915 | SEQ ID NO 1215 | Contig53226_RC | SEQ ID NO 2525 |
| | NM_006096 | SEQ ID NO 1240 | Contig53268_RC | SEQ ID NO 2529 |
| | NM_006101 | SEQ ID NO 1241 | Contig53646_RC | SEQ ID NO 2538 |
| | NM_006115 | SEQ ID NO 1245 | Contig53742_RC | SEQ ID NO 2542 |
| | NM_006117 | SEQ ID NO 1246 | Contig55188_RC | SEQ ID NO 2586 |
| 30 | NM_006201 | SEQ ID NO 1254 | Contig55313_RC | SEQ ID NO 2590 |
| | NM_006265 | SEQ ID NO 1260 | Contig55377_RC | SEQ ID NO 2591 |
| | NM_006281 | SEQ ID NO 1263 | Contig55725_RC | SEQ ID NO 2600 |
| | NM_006372 | SEQ ID NO 1273 | Contig55813_RC | SEQ ID NO 2607 |
| | NM_006681 | SEQ ID NO 1306 | Contig55829_RC | SEQ ID NO 2608 |
| 35 | NM_006763 | SEQ ID NO 1315 | Contig56457_RC | SEQ ID NO 2630 |
| | NM_006931 | SEQ ID NO 1341 | Contig57595 | SEQ ID NO 2655 |

| GenBank Accession Number | SEQ ID NO | GenBank Accession Number | SEQ ID NO |
|-----------------------------|----------------|-----------------------------|----------------|
| NM_007036 | SEQ ID NO 1349 | Contig57864_RC | SEQ ID NO 2663 |
| NM_007203 | SEQ ID NO 1362 | Contig58368_RC | SEQ ID NO 2668 |
| NM_012177 | SEQ ID NO 1390 | Contig60864_RC | SEQ ID NO 2676 |
| NM_012214 | SEQ ID NO 1392 | Contig63102_RC | SEQ ID NO 2684 |
| NM_012261 | SEQ ID NO 1397 | Contig63649_RC | SEQ ID NO 2686 |
| NM_012429 | SEQ ID NO 1413 | Contig64688 | SEQ ID NO 2690 |
| NM_013262 | SEQ ID NO 1425 | | |

5

10

15

20

25

30

35

Table 6. 70 Preferred prognosis markers drawn from Table 5.

| Identifier | Correlation | Sequence Name | Description |
|-------------------|-------------|---------------|--|
| AL080059 | -0.527150 | | Homo sapiens mRNA for KIAA1750 protein, partial cds |
| 5 Contig63649_RC | -0.468130 | | ESTs |
| Contig46218_RC | -0.432540 | | ESTs |
| NM_016359 | -0.424930 | LOC51203 | clone HQ0310 PRO0310p1 |
| 10 AA555029_RC | -0.424120 | | ESTs |
| NM_003748 | 0.420671 | ALDH4 | aldehyde dehydrogenase 4 (glutamate gamma-semialdehyde dehydrogenase; pyrroline-5-carboxylate dehydrogenase) |
| Contig38288_RC | -0.414970 | | ESTs, Weakly similar to ISHUSS protein disulfide-isomerase [H.sapiens] |
| 15 NM_003862 | 0.410964 | FGF18 | fibroblast growth factor 18 |
| Contig28552_RC | -0.409260 | | Homo sapiens mRNA; cDNA DKFZp434C0931 (from clone DKFZp434C0931); partial cds |
| Contig32125_RC | 0.409054 | | ESTs |
| 20 U82987 | 0.407002 | BBC3 | Bcl-2 binding component 3 |
| AL137718 | -0.404980 | | Homo sapiens mRNA; cDNA DKFZp434C0931 (from clone DKFZp434C0931); partial cds |
| AB037863 | 0.402335 | KIAA1442 | KIAA1442 protein |
| NM_020188 | -0.400070 | DC13 | DC13 protein |
| 25 NM_020974 | 0.399987 | CEGP1 | CEGP1 protein |
| NM_000127 | -0.399520 | EXT1 | exostoses (multiple) 1 |
| NM_002019 | -0.398070 | FLT1 | fms-related tyrosine kinase 1 (vascular endothelial growth factor/vascular permeability factor receptor) |
| 30 NM_002073 | -0.395460 | GNAZ | guanine nucleotide binding protein (G protein), alpha z polypeptide |
| NM_000436 | -0.392120 | OXCT | 3-oxoacid CoA transferase |
| NM_004994 | -0.391690 | MMP9 | matrix metalloproteinase 9 (gelatinase B, 92kD gelatinase, 92kD type IV collagenase) |
| 35 Contig55377_RC | 0.390600 | | ESTs |

| | | | | |
|----|--------------------|-----------|-------------------|---|
| | Contig35251_ RC | -0.390410 | | Homo sapiens cDNA: FLJ22719 fis, clone HSI14307 |
| | Contig25991 | -0.390370 | ECT2 | epithelial cell transforming sequence 2 oncogene |
| | NM_003875 | -0.386520 | GMPS | guanine monphosphate synthetase |
| 5 | NM_006101 | -0.385890 | HEC | highly expressed in cancer, rich in leucine heptad repeats |
| | NM_003882 | 0.384479 | WISP1 | WNT1 inducible signaling pathway protein 1 |
| | NM_003607 | -0.384390 | PK428 | Ser-Thr protein kinase related to the myotonic dystrophy protein kinase |
| 10 | AF073519 | -0.383340 | SERF1A | small EDRK-rich factor 1A (telomeric) |
| | AF052162 | -0.380830 | FLJ12443 | hypothetical protein FLJ12443 |
| | NM_000849 | 0.380831 | GSTM3 | glutathione S-transferase M3 (brain) |
| | Contig32185_ RC | -0.379170 | | Homo sapiens cDNA FLJ13997 fis, clone Y79AA1002220 |
| 15 | NM_016577 | -0.376230 | RAB6B | RAB6B, member RAS oncogene family |
| | Contig48328_ RC | 0.375252 | | ESTs, Weakly similar to T17248 hypothetical protein DKFZp586G1122.1 [H.sapiens] |
| | Contig46223_ RC | 0.374289 | | ESTs |
| 20 | NM_015984 | -0.373880 | UCH37 | ubiquitin C-terminal hydrolase UCH37 |
| | NM_006117 | 0.373290 | PECI | peroxisomal D3,D2-enoyl-CoA isomerase |
| | AK000745 | -0.373060 | | Homo sapiens cDNA FLJ20738 fis, clone HEP08257 |
| 25 | Contig40831_ RC | -0.372930 | | ESTs |
| | NM_003239 | 0.371524 | TGFB3 | transforming growth factor, beta 3 |
| | NM_014791 | -0.370860 | KIAA0175 | KIAA0175 gene product |
| | X05610 | -0.370860 | COL4A2 | collagen, type IV, alpha 2 |
| | NM_016448 | -0.369420 | L2DTL | L2DTL protein |
| 30 | NM_018401 | 0.368349 | HSA250839 | gene for serine/threonine protein kinase |
| | NM_000788 | -0.367700 | DCK | deoxycytidine kinase |
| | Contig51464_ RC | -0.367450 | FLJ22477 | hypothetical protein FLJ22477 |
| 35 | AL080079 | -0.367390 | DKFZP564D 0462 | hypothetical protein DKFZp564D0462 |

| | | | | |
|----|----------------|-----------|----------|--|
| | NM_006931 | -0.366490 | SLC2A3 | solute carrier family 2 (facilitated glucose transporter), member 3 |
| | AF257175 | 0.365900 | | Homo sapiens hepatocellular carcinoma-associated antigen 64 (HCA64) mRNA, complete cds |
| 5 | NM_014321 | -0.365810 | ORC6L | origin recognition complex, subunit 6 (yeast homolog)-like |
| | NM_002916 | -0.365590 | RFC4 | replication factor C (activator 1) 4 (37kD) |
| | Contig55725_RC | -0.365350 | | ESTs; Moderately similar to T50635 hypothetical protein DKFZp762L0311.1 [H.sapiens] |
| 10 | Contig24252_RC | -0.364990 | | ESTs |
| | AF201951 | 0.363953 | CFFM4 | high affinity immunoglobulin epsilon receptor beta subunit |
| | NM_005915 | -0.363850 | MCM6 | minichromosome maintenance deficient (mis5, S. pombe) 6 |
| 15 | NM_001282 | 0.363326 | AP2B1 | adaptor-related protein complex 2, beta 1 subunit |
| | Contig56457_RC | -0.361650 | TMEFF1 | transmembrane protein with EGF-like and two follistatin-like domains 1 |
| | NM_000599 | -0.361290 | IGFBP5 | insulin-like growth factor binding protein 5 |
| 20 | NM_020386 | -0.360780 | LOC57110 | H-REV107 protein-related protein |
| | NM_014889 | -0.360040 | MP1 | metalloprotease 1 (pitrilysin family) |
| | AF055033 | -0.359940 | IGFBP5 | insulin-like growth factor binding protein 5 |
| | NM_006681 | -0.359700 | NMU | neuromedin U |
| | NM_007203 | -0.359570 | AKAP2 | A kinase (PRKA) anchor protein 2 |
| 25 | Contig63102_RC | 0.359255 | FLJ11354 | hypothetical protein FLJ11354 |
| | NM_003981 | -0.358260 | PRC1 | protein regulator of cytokinesis 1 |
| | Contig20217_RC | -0.357880 | | ESTs |
| | NM_001809 | -0.357720 | CENPA | centromere protein A (17kD) |
| 30 | Contig2399_RC | -0.356600 | SM-20 | similar to rat smooth muscle protein SM-20 |
| | NM_004702 | -0.356600 | CCNE2 | cyclin E2 |
| | NM_007036 | -0.356540 | ESM1 | endothelial cell-specific molecule 1 |
| | NM_018354 | -0.356000 | FLJ11190 | hypothetical protein FLJ11190 |

The sets of markers listed in Tables 1-6 partially overlap; in other words, some markers are present in multiple sets, while other markers are unique to a set (FIG. 1). Thus, in one embodiment, the invention provides a set of 256 genetic markers that can distinguish between ER(+) and ER(-), and also between *BRCA1* tumors and sporadic tumors
5 (*i.e.*, classify a tumor as ER(-) or ER(-) and *BRCA1*-related or sporadic). In a more specific embodiment, the invention provides subsets of at least 20, at least 50, at least 100, or at least 150 of the set of 256 markers, that can classify a tumor as ER(-) or ER(-) and *BRCA1*-related or sporadic. In another embodiment, the invention provides 165 markers that can distinguish between ER(+) and ER(-), and also between patients with good versus poor
10 prognosis (*i.e.*, classify a tumor as either ER(-) or ER(+) and as having been removed from a patient with a good prognosis or a poor prognosis). In a more specific embodiment, the invention further provides subsets of at least 20, 50, 100 or 125 of the full set of 165 markers, which also classify a tumor as either ER(-) or ER(+) and as having been removed from a patient with a good prognosis or a poor prognosis. The invention further provides a
15 set of twelve markers that can distinguish between *BRCA1* tumors and sporadic tumors, and between patients with good versus poor prognosis. Finally, the invention provides eleven markers capable of differentiating all three statuses. Conversely, the invention provides 2,050 of the 2,460 ER-status markers that can determine only ER status, 173 of the 430 *BRCA1* v. sporadic markers that can determine only *BRCA1* v. sporadic status, and 65 of the
20 231 prognosis markers that can only determine prognosis. In more specific embodiments, the invention also provides for subsets of at least 20, 50, 100, 200, 500, 1,000, 1,500 or 2,000 of the 2,050 ER-status markers that also determine only ER status. The invention also provides subsets of at least 20, 50, 100 or 150 of the 173 markers that also determine only *BRCA1* v. sporadic status. The invention further provides subsets of at least 20, 30, 40,
25 or 50 of the 65 prognostic markers that also determine only prognostic status.

Any of the sets of markers provided above may be used alone specifically or in combination with markers outside the set. For example, markers that distinguish ER-status may be used in combination with the *BRCA1* vs. sporadic markers, or with the prognostic markers, or both. Any of the marker sets provided above may also be used in
30 combination with other markers for breast cancer, or for any other clinical or physiological condition.

The relationship between the marker sets is diagramed in FIG. 1.

5.3.2 IDENTIFICATION OF MARKERS

The present invention provides sets of markers for the identification of conditions or indications associated with breast cancer. Generally, the marker sets were identified by determining which of ~25,000 human markers had expression patterns that
5 correlated with the conditions or indications.

In one embodiment, the method for identifying marker sets is as follows. After extraction and labeling of target polynucleotides, the expression of all markers (genes) in a sample X is compared to the expression of all markers in a standard or control. In one embodiment, the standard or control comprises target polynucleotide molecules derived
10 from a sample from a normal individual (*i.e.*, an individual not afflicted with breast cancer). In a preferred embodiment, the standard or control is a pool of target polynucleotide molecules. The pool may be derived from collected samples from a number of normal individuals. In a preferred embodiment, the pool comprises samples taken from a number of individuals having sporadic-type tumors. In another preferred embodiment, the pool
15 comprises an artificially-generated population of nucleic acids designed to approximate the level of nucleic acid derived from each marker found in a pool of marker-derived nucleic acids derived from tumor samples. In yet another embodiment, the pool is derived from normal or breast cancer cell lines or cell line samples.

The comparison may be accomplished by any means known in the art. For
20 example, expression levels of various markers may be assessed by separation of target polynucleotide molecules (*e.g.*, RNA or cDNA) derived from the markers in agarose or polyacrylamide gels, followed by hybridization with marker-specific oligonucleotide probes. Alternatively, the comparison may be accomplished by the labeling of target polynucleotide molecules followed by separation on a sequencing gel. Polynucleotide
25 samples are placed on the gel such that patient and control or standard polynucleotides are in adjacent lanes. Comparison of expression levels is accomplished visually or by means of densitometer. In a preferred embodiment, the expression of all markers is assessed simultaneously by hybridization to a microarray. In each approach, markers meeting certain criteria are identified as associated with breast cancer.

30 A marker is selected based upon significant difference of expression in a sample as compared to a standard or control condition. Selection may be made based upon either significant up- or down regulation of the marker in the patient sample. Selection may also be made by calculation of the statistical significance (*i.e.*, the p-value) of the correlation between the expression of the marker and the condition or indication. Preferably, both
35 selection criteria are used. Thus, in one embodiment of the present invention, markers

associated with breast cancer are selected where the markers show both more than two-fold change (increase or decrease) in expression as compared to a standard, and the p-value for the correlation between the existence of breast cancer and the change in marker expression is no more than 0.01 (*i.e.*, is statistically significant).

- 5 The expression of the identified breast cancer-related markers is then used to identify markers that can differentiate tumors into clinical types. In a specific embodiment using a number of tumor samples, markers are identified by calculation of correlation coefficients between the clinical category or clinical parameter(s) and the linear, logarithmic or any transform of the expression ratio across all samples for each individual gene.
- 10 Specifically, the correlation coefficient is calculated as

$$\rho = (\vec{c} \bullet \vec{r}) / (\|\vec{c}\| \cdot \|\vec{r}\|)$$

Equation (2)

- where \vec{c} represents the clinical parameters or categories and \vec{r} represents the linear, logarithmic or any transform of the ratio of expression between sample and control.
- 15 Markers for which the coefficient of correlation exceeds a cutoff are identified as breast cancer-related markers specific for a particular clinical type. Such a cutoff or threshold corresponds to a certain significance of discriminating genes obtained by Monte Carlo simulations. The threshold depends upon the number of samples used; the threshold can be
- 20 calculated as $3 \times 1/\sqrt{n-3}$, where $1/\sqrt{n-3}$ is the distribution width and n = the number of samples. In a specific embodiment, markers are chosen if the correlation coefficient is greater than about 0.3 or less than about -0.3.

- Next, the significance of the correlation is calculated. This significance may be calculated by any statistical means by which such significance is calculated. In a specific
- 25 example, a set of correlation data is generated using a Monte-Carlo technique to randomize the association between the expression difference of a particular marker and the clinical category. The frequency distribution of markers satisfying the criteria through calculation of correlation coefficients is compared to the number of markers satisfying the criteria in the data generated through the Monte-Carlo technique. The frequency distribution of markers
- 30 satisfying the criteria in the Monte-Carlo runs is used to determine whether the number of markers selected by correlation with clinical data is significant. See Example 4.

- Once a marker set is identified, the markers may be rank-ordered in order of significance of discrimination. One means of rank ordering is by the amplitude of correlation between the change in gene expression of the marker and the specific condition
- 35 being discriminated. Another, preferred means is to use a statistical metric. In a specific embodiment, the metric is a Fisher-like statistic:

$$t = \frac{(\langle x_1 \rangle - \langle x_2 \rangle)}{\sqrt{[\sigma_1^2(n_1 - 1) + \sigma_2^2(n_2 - 1)] / (n_1 + n_2 - 1) / (1/n_1 + 1/n_2)}} \quad \text{Equation (3)}$$

- 5 In this equation, $\langle x_1 \rangle$ is the error-weighted average of the log ratio of transcript expression measurements within a first diagnostic group (e.g., ER(-)), $\langle x_2 \rangle$ is the error-weighted average of log ratio within a second, related diagnostic group (e.g., ER(+)), σ_1 is the variance of the
- 10 log ratio within the ER(-) group and n_1 is the number of samples for which valid measurements of log ratios are available. σ_2 is the variance of log ratio within the second diagnostic group (e.g., ER(+)), and n_2 is the number of samples for which valid measurements of log ratios are available. The t -value represents the variance-compensated
- 15 difference between two means.

The rank-ordered marker set may be used to optimize the number of markers in the set used for discrimination. This is accomplished generally in a "leave one out" method as follows. In a first run, a subset, for example 5, of the markers from the top of the ranked list is used to generate a template, where out of X samples, $X-1$ are used to generate

20 the template, and the status of the remaining sample is predicted. This process is repeated for every sample until every one of the X samples is predicted once. In a second run, additional markers, for example 5, are added, so that a template is now generated from 10 markers, and the outcome of the remaining sample is predicted. This process is repeated until the entire set of markers is used to generate the template. For each of the runs, type 1

25 error (false negative) and type 2 errors (false positive) are counted; the optimal number of markers is that number where the type 1 error rate, or type 2 error rate, or preferably the total of type 1 and type 2 error rate is lowest.

For prognostic markers, validation of the marker set may be accomplished by an additional statistic, a survival model. This statistic generates the probability of tumor

30 distant metastases as a function of time since initial diagnosis. A number of models may be used, including Weibull, normal, log-normal, log logistic, log-exponential, or log-Rayleigh (Chapter 12 "Life Testing", S-PLUS 2000 GUIDE TO STATISTICS, Vol. 2, p. 368 (2000)). For the "normal" model, the probability of distant metastases P at time t is calculated as

$$35 \quad P = \alpha \times \exp(-t^2/\tau^2) \quad \text{Equation (4)}$$

where α is fixed and equal to 1, and τ is a parameter to be fitted and measures the "expected lifetime".

It will be apparent to those skilled in the art that the above methods, in particular the statistical methods, described above, are not limited to the identification of markers associated with breast cancer, but may be used to identify set of marker genes associated with any phenotype. The phenotype can be the presence or absence of a disease such as cancer, or the presence or absence of any identifying clinical condition associated with that cancer. In the disease context, the phenotype may be a prognosis such as a survival time, probability of distant metastases of a disease condition, or likelihood of a particular response to a therapeutic or prophylactic regimen. The phenotype need not be cancer, or a disease; the phenotype may be a nominal characteristic associated with a healthy individual.

5.3.3 SAMPLE COLLECTION

In the present invention, target polynucleotide molecules are extracted from a sample taken from an individual afflicted with breast cancer. The sample may be collected in any clinically acceptable manner, but must be collected such that marker-derived polynucleotides (*i.e.*, RNA) are preserved. mRNA or nucleic acids derived therefrom (*i.e.*, cDNA or amplified DNA) are preferably labeled distinguishably from standard or control polynucleotide molecules, and both are simultaneously or independently hybridized to a microarray comprising some or all of the markers or marker sets or subsets described above. Alternatively, mRNA or nucleic acids derived therefrom may be labeled with the same label as the standard or control polynucleotide molecules, wherein the intensity of hybridization of each at a particular probe is compared. A sample may comprise any clinically relevant tissue sample, such as a tumor biopsy or fine needle aspirate, or a sample of bodily fluid, such as blood, plasma, serum, lymph, ascitic fluid, cystic fluid, urine or nipple exudate. The sample may be taken from a human, or, in a veterinary context, from non-human animals such as ruminants, horses, swine or sheep, or from domestic companion animals such as felines and canines.

Methods for preparing total and poly(A)+ RNA are well known and are described generally in Sambrook *et al.*, MOLECULAR CLONING - A LABORATORY MANUAL (2ND ED.), Vols. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York (1989)) and Ausubel *et al.*, CURRENT PROTOCOLS IN MOLECULAR BIOLOGY, vol. 2, Current Protocols Publishing, New York (1994)).

RNA may be isolated from eukaryotic cells by procedures that involve lysis of the cells and denaturation of the proteins contained therein. Cells of interest include wild-type cells (*i.e.*, non-cancerous), drug-exposed wild-type cells, tumor- or tumor-derived cells, modified cells, normal or tumor cell line cells, and drug-exposed modified cells.

5 Additional steps may be employed to remove DNA. Cell lysis may be accomplished with a nonionic detergent, followed by microcentrifugation to remove the nuclei and hence the bulk of the cellular DNA. In one embodiment, RNA is extracted from cells of the various types of interest using guanidinium thiocyanate lysis followed by CsCl centrifugation to separate the RNA from DNA (Chirgwin *et al.*, *Biochemistry* 18:5294-5299
10 (1979)). Poly(A)+ RNA is selected by selection with oligo-dT cellulose (*see* Sambrook *et al.*, *MOLECULAR CLONING - A LABORATORY MANUAL* (2ND ED.), Vols. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York (1989). Alternatively, separation of RNA from DNA can be accomplished by organic extraction, for example, with hot phenol or phenol/chloroform/isoamyl alcohol.

15 If desired, RNase inhibitors may be added to the lysis buffer. Likewise, for certain cell types, it may be desirable to add a protein denaturation/digestion step to the protocol.

For many applications, it is desirable to preferentially enrich mRNA with respect to other cellular RNAs, such as transfer RNA (tRNA) and ribosomal RNA (rRNA).
20 Most mRNAs contain a poly(A) tail at their 3' end. This allows them to be enriched by affinity chromatography, for example, using oligo(dT) or poly(U) coupled to a solid support, such as cellulose or Sephadex™ (*see* Ausubel *et al.*, *CURRENT PROTOCOLS IN MOLECULAR BIOLOGY*, vol. 2, Current Protocols Publishing, New York (1994). Once bound, poly(A)+ mRNA is eluted from the affinity column using 2 mM EDTA/0.1% SDS.

25 The sample of RNA can comprise a plurality of different mRNA molecules, each different mRNA molecule having a different nucleotide sequence. In a specific embodiment, the mRNA molecules in the RNA sample comprise at least 100 different nucleotide sequences. More preferably, the mRNA molecules of the RNA sample comprise mRNA molecules corresponding to each of the marker genes. In another specific
30 embodiment, the RNA sample is a mammalian RNA sample.

In a specific embodiment, total RNA or mRNA from cells are used in the methods of the invention. The source of the RNA can be cells of a plant or animal, human, mammal, primate, non-human animal, dog, cat, mouse, rat, bird, yeast, eukaryote, prokaryote, etc. In specific embodiments, the method of the invention is used with a sample
35 containing total mRNA or total RNA from 1×10^6 cells or less. In another embodiment,

proteins can be isolated from the foregoing sources, by methods known in the art, for use in expression analysis at the protein level.

Probes to the homologs of the marker sequences disclosed herein can be employed preferably wherein non-human nucleic acid is being assayed.

5

5.4 METHODS OF USING BREAST CANCER MARKER SETS

5.4.1 DIAGNOSTIC METHODS

The present invention provides for methods of using the marker sets to analyze a sample from an individual so as to determine the individual's tumor type or subtype at a molecular level, whether a tumor is of the ER(+) or ER(-) type, and whether the tumor is *BRCA1*-associated or sporadic. The individual need not actually be afflicted with breast cancer. Essentially, the expression of specific marker genes in the individual, or a sample taken therefrom, is compared to a standard or control. For example, assume two breast cancer-related conditions, X and Y. One can compare the level of expression of breast cancer prognostic markers for condition X in an individual to the level of the marker-derived polynucleotides in a control, wherein the level represents the level of expression exhibited by samples having condition X. In this instance, if the expression of the markers in the individual's sample is substantially (*i.e.*, statistically) different from that of the control, then the individual does not have condition X. Where, as here, the choice is bimodal (*i.e.*, a sample is either X or Y), the individual can additionally be said to have condition Y. Of course, the comparison to a control representing condition Y can also be performed. Preferably both are performed simultaneously, such that each control acts as both a positive and a negative control. The distinguishing result may thus either be a demonstrable difference from the expression levels (*i.e.*, the amount of marker-derived RNA, or polynucleotides derived therefrom) represented by the control, or no significant difference.

Thus, in one embodiment, the method of determining a particular tumor-related status of an individual comprises the steps of (1) hybridizing labeled target polynucleotides from an individual to a microarray containing one of the above marker sets; (2) hybridizing standard or control polynucleotides molecules to the microarray, wherein the standard or control molecules are differentially labeled from the target molecules; and (3) determining the difference in transcript levels, or lack thereof, between the target and standard or control, wherein the difference, or lack thereof, determines the individual's tumor-related status. In a more specific embodiment, the standard or control molecules comprise marker-derived polynucleotides from a pool of samples from normal individuals,

or a pool of tumor samples from individuals having sporadic-type tumors. In a preferred embodiment, the standard or control is an artificially-generated pool of marker-derived polynucleotides, which pool is designed to mimic the level of marker expression exhibited by clinical samples of normal or breast cancer tumor tissue having a particular clinical indication (*i.e.*, cancerous or non-cancerous; ER(+) or ER(-) tumor; *BRCA1*- or sporadic type tumor). In another specific embodiment, the control molecules comprise a pool derived from normal or breast cancer cell lines.

The present invention provides sets of markers useful for distinguishing ER(+) from ER(-) tumor types. Thus, in one embodiment of the above method, the level of polynucleotides (*i.e.*, mRNA or polynucleotides derived therefrom) in a sample from an individual, expressed from the markers provided in Table 1 are compared to the level of expression of the same markers from a control, wherein the control comprises marker-related polynucleotides derived from ER(+) samples, ER(-) samples, or both. Preferably, the comparison is to both ER(+) and ER(-), and preferably the comparison is to polynucleotide pools from a number of ER(+) and ER(-) samples, respectively. Where the individual's marker expression most closely resembles or correlates with the ER(+) control, and does not resemble or correlate with the ER(-) control, the individual is classified as ER(+). Where the pool is not pure ER(+) or ER(-), for example, a sporadic pool is used. A set of experiments using individuals with known ER status should be hybridized against the pool, in order to define the expression templates for the ER(+) and ER(-) group. Each individual with unknown ER status is hybridized against the same pool and the expression profile is compared to the templates (s) to determine the individual's ER status.

The present invention provides sets of markers useful for distinguishing *BRCA1*-related tumors from sporadic tumors. Thus, the method can be performed substantially as for the ER(+/-) determination, with the exception that the markers are those listed in Tables 3 and 4, and the control markers are a pool of marker-derived polynucleotides *BRCA1* tumor samples, and a pool of marker-derived polynucleotides from sporadic tumors. A patient is determined to have a *BRCA1* germline mutation where the expression of the individual's marker-derived polynucleotides most closely resemble, or are most closely correlated with, that of the *BRCA1* control. Where the control is not pure *BRCA1* or sporadic, two templates can be defined in a manner similar to that for ER status, as described above.

For the above two embodiments of the method, the full set of markers may be used (*i.e.*, the complete set of markers for Tables 1 or 3). In other embodiments, subsets

35

of the markers may be used. In a preferred embodiment, the preferred markers listed in Tables 2 or 4 are used.

The similarity between the marker expression profile of an individual and that of a control can be assessed a number of ways. In the simplest case, the profiles can be compared visually in a printout of expression difference data. Alternatively, the similarity can be calculated mathematically.

In one embodiment, the similarity measure between two patients x and y, or patient x and a template y, can be calculated using the following equation:

$$S = 1 - \left[\frac{\sum_{i=1}^{N_p} \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sigma_{x_i} \sigma_{y_i}}}{\sqrt{\sum_{i=1}^{N_p} \left(\frac{x_i - \bar{x}}{\sigma_{x_i}} \right)^2 \sum_{i=1}^{N_p} \left(\frac{y_i - \bar{y}}{\sigma_{y_i}} \right)^2}} \right] \quad \text{Equation (5)}$$

In this equation, x and y are two patients with components of log ratio x_i and y_i ,

$i=1, \dots, N=4,986$. Associated with every value x_i is error σ_{x_i} . The smaller the value σ_{x_i} ,

the more reliable the measurement x_i . $\bar{x} = \frac{\sum_{i=1}^{N_p} x_i}{\sum_{i=1}^{N_p} \frac{1}{\sigma_{x_i}^2}}$ is the error-weighted arithmetic mean.

In a preferred embodiment, templates are developed for sample comparison. The template is defined as the error-weighted log ratio average of the expression difference for the group of marker genes able to differentiate the particular breast cancer-related condition. For example, templates are defined for ER(+) samples and for ER(-) samples. Next, a classifier parameter is calculated. This parameter may be calculated using either expression level differences between the sample and template, or by calculation of a correlation coefficient. Such a coefficient, P_i , can be calculated using the following equation:

$$P_i = (\bar{z}_i \bullet \bar{y}) / (\|\bar{z}_i\| \cdot \|\bar{y}\|) \quad \text{Equation (1)}$$

where Z_i is the expression template i , and y is the expression profile of a patient.

Thus, in a more specific embodiment, the above method of determining a particular tumor-related status of an individual comprises the steps of (1) hybridizing

labeled target polynucleotides from an individual to a microarray containing one of the above marker sets; (2) hybridizing standard or control polynucleotides molecules to the microarray, wherein the standard or control molecules are differentially labeled from the target molecules; and (3) determining the ratio (or difference) of transcript levels between
5 two channels (individual and control), or simply the transcript levels of the individual; and (4) comparing the results from (3) to the predefined templates, wherein said determining is accomplished by means of the statistic of Equation 1 or Equation 5, and wherein the difference, or lack thereof, determines the individual's tumor-related status.

10

5.4.2 PROGNOSTIC METHODS

The present invention provides sets of markers useful for distinguishing samples from those patients with a good prognosis from samples from patients with a poor prognosis. Thus, the invention further provides a method for using these markers to determine whether an individual afflicted with breast cancer will have a good or poor
15 clinical prognosis. In one embodiment, the invention provides for method of determining whether an individual afflicted with breast cancer will likely experience a relapse within five years of initial diagnosis (*i.e.*, whether an individual has a poor prognosis) comprising (1) comparing the level of expression of the markers listed in Table 5 in a sample taken from the individual to the level of the same markers in a standard or control, where the
20 standard or control levels represent those found in an individual with a poor prognosis; and (2) determining whether the level of the marker-related polynucleotides in the sample from the individual is significantly different than that of the control, wherein if no substantial difference is found, the patient has a poor prognosis, and if a substantial difference is found, the patient has a good prognosis. Persons of skill in the art will readily see that the markers
25 associated with good prognosis can also be used as controls. In a more specific embodiment, both controls are run. In case the pool is not pure 'good prognosis' or 'poor prognosis', a set of experiments of individuals with known outcome should be hybridized against the pool to define the expression templates for the good prognosis and poor prognosis group. Each individual with unknown outcome is hybridized against the same
30 pool and the resulting expression profile is compared to the templates to predict its outcome.

Poor prognosis of breast cancer may indicate that a tumor is relatively aggressive, while good prognosis may indicate that a tumor is relatively nonaggressive. Therefore, the invention provides for a method of determining a course of treatment of a
35 breast cancer patient, comprising determining whether the level of expression of the 231

markers of Table 5, or a subset thereof, correlates with the level of these markers in a sample representing a good prognosis expression pattern or a poor prognosis pattern; and determining a course of treatment, wherein if the expression correlates with the poor prognosis pattern, the tumor is treated as an aggressive tumor.

5 As with the diagnostic markers, the method can use the complete set of markers listed in Table 5. However, subsets of the markers may also be used. In a preferred embodiment, the subset listed in Table 6 is used.

 Classification of a sample as "good prognosis" or "poor prognosis" is accomplished substantially as for the diagnostic markers described above, wherein a
10 template is generated to which the marker expression levels in the sample are compared.

 The use of marker sets is not restricted to the prognosis of breast cancer-related conditions, and may be applied in a variety of phenotypes or conditions, clinical or experimental, in which gene expression plays a role. Where a set of markers has been identified that corresponds to two or more phenotypes, the marker sets can be used to
15 distinguish these phenotypes. For example, the phenotypes may be the diagnosis and/or prognosis of clinical states or phenotypes associated with other cancers, other disease conditions, or other physiological conditions, wherein the expression level data is derived from a set of genes correlated with the particular physiological or disease condition.

20 5.4.3 IMPROVING SENSITIVITY TO EXPRESSION LEVEL DIFFERENCES

 In using the markers disclosed herein, and, indeed, using any sets of markers to differentiate an individual having one phenotype from another individual having a second phenotype, one can compare the absolute expression of each of the markers in a sample to a control; for example, the control can be the average level of expression of each of the
25 markers, respectively, in a pool of individuals. To increase the sensitivity of the comparison, however, the expression level values are preferably transformed in a number of ways.

 For example, the expression level of each of the markers can be normalized by the average expression level of all markers the expression level of which is determined,
30 or by the average expression level of a set of control genes. Thus, in one embodiment, the markers are represented by probes on a microarray, and the expression level of each of the markers is normalized by the mean or median expression level across all of the genes represented on the microarray, including any non-marker genes. In a specific embodiment, the normalization is carried out by dividing the median or mean level of expression of all of
35 the genes on the microarray. In another embodiment, the expression levels of the markers is

normalized by the mean or median level of expression of a set of control markers. In a specific embodiment, the control markers comprise a set of housekeeping genes. In another specific embodiment, the normalization is accomplished by dividing by the median or mean expression level of the control genes.

5 The sensitivity of a marker-based assay will also be increased if the expression levels of individual markers are compared to the expression of the same markers in a pool of samples. Preferably, the comparison is to the mean or median expression level of each the marker genes in the pool of samples. Such a comparison may be accomplished, for example, by dividing by the mean or median expression level of the pool for each of the
10 markers from the expression level each of the markers in the sample. This has the effect of accentuating the relative differences in expression between markers in the sample and markers in the pool as a whole, making comparisons more sensitive and more likely to produce meaningful results than the use of absolute expression levels alone. The expression level data may be transformed in any convenient way; preferably, the expression level data
15 for all is log transformed before means or medians are taken.

 In performing comparisons to a pool, two approaches may be used. First, the expression levels of the markers in the sample may be compared to the expression level of those markers in the pool, where nucleic acid derived from the sample and nucleic acid derived from the pool are hybridized during the course of a single experiment. Such an
20 approach requires that new pool nucleic acid be generated for each comparison or limited numbers of comparisons, and is therefore limited by the amount of nucleic acid available. Alternatively, and preferably, the expression levels in a pool, whether normalized and/or transformed or not, are stored on a computer, or on computer-readable media, to be used in comparisons to the individual expression level data from the sample (*i.e.*, single-channel
25 data).

 Thus, the current invention provides the following method of classifying a first cell or organism as having one of at least two different phenotypes, where the different phenotypes comprise a first phenotype and a second phenotype. The level of expression of each of a plurality of genes in a first sample from the first cell or organism is compared to
30 the level of expression of each of said genes, respectively, in a pooled sample from a plurality of cells or organisms, the plurality of cells or organisms comprising different cells or organisms exhibiting said at least two different phenotypes, respectively, to produce a first compared value. The first compared value is then compared to a second compared value, wherein said second compared value is the product of a method comprising
35 comparing the level of expression of each of said genes in a sample from a cell or organism

characterized as having said first phenotype to the level of expression of each of said genes, respectively, in the pooled sample. The first compared value is then compared to a third compared value, wherein said third compared value is the product of a method comprising comparing the level of expression of each of the genes in a sample from a cell or organism
5 characterized as having the second phenotype to the level of expression of each of the genes, respectively, in the pooled sample. Optionally, the first compared value can be compared to additional compared values, respectively, where each additional compared value is the product of a method comprising comparing the level of expression of each of said genes in a sample from a cell or organism characterized as having a phenotype different
10 from said first and second phenotypes but included among the at least two different phenotypes, to the level of expression of each of said genes, respectively, in said pooled sample. Finally, a determination is made as to which of said second, third, and, if present, one or more additional compared values, said first compared value is most similar, wherein the first cell or organism is determined to have the phenotype of the cell or organism used to
15 produce said compared value most similar to said first compared value.

In a specific embodiment of this method, the compared values are each ratios of the levels of expression of each of said genes. In another specific embodiment, each of the levels of expression of each of the genes in the pooled sample are normalized prior to any of the comparing steps. In a more specific embodiment, the normalization of the levels
20 of expression is carried out by dividing by the median or mean level of the expression of each of the genes or dividing by the mean or median level of expression of one or more housekeeping genes in the pooled sample from said cell or organism. In another specific embodiment, the normalized levels of expression are subjected to a log transform, and the comparing steps comprise subtracting the log transform from the log of the levels of
25 expression of each of the genes in the sample. In another specific embodiment, the two or more different phenotypes are different stages of a disease or disorder. In still another specific embodiment, the two or more different phenotypes are different prognoses of a disease or disorder. In yet another specific embodiment, the levels of expression of each of the genes, respectively, in the pooled sample or said levels of expression of each of said
30 genes in a sample from the cell or organism characterized as having the first phenotype, second phenotype, or said phenotype different from said first and second phenotypes, respectively, are stored on a computer or on a computer-readable medium.

In another specific embodiment, the two phenotypes are ER(+) or ER(-) status. In another specific embodiment, the two phenotypes are *BRCA1* or sporadic tumor-
35

type status. In yet another specific embodiment, the two phenotypes are good prognosis and poor prognosis.

Of course, single-channel data may also be used without specific comparison to a mathematical sample pool. For example, a sample may be classified as having a first or a second phenotype, wherein the first and second phenotypes are related, by calculating the similarity between the expression of at least 5 markers in the sample, where the markers are correlated with the first or second phenotype, to the expression of the same markers in a first phenotype template and a second phenotype template, by (a) labeling nucleic acids derived from a sample with a fluorophore to obtain a pool of fluorophore-labeled nucleic acids; (b) contacting said fluorophore-labeled nucleic acid with a microarray under conditions such that hybridization can occur, detecting at each of a plurality of discrete loci on the microarray a fluorescent emission signal from said fluorophore-labeled nucleic acid that is bound to said microarray under said conditions; and (c) determining the similarity of marker gene expression in the individual sample to the first and second templates, wherein if said expression is more similar to the first template, the sample is classified as having the first phenotype, and if said expression is more similar to the second template, the sample is classified as having the second phenotype.

5.5 DETERMINATION OF MARKER GENE EXPRESSION LEVELS

5.5.1 METHODS

The expression levels of the marker genes in a sample may be determined by any means known in the art. The expression level may be determined by isolating and determining the level (*i.e.*, amount) of nucleic acid transcribed from each marker gene. Alternatively, or additionally, the level of specific proteins translated from mRNA transcribed from a marker gene may be determined.

The level of expression of specific marker genes can be accomplished by determining the amount of mRNA, or polynucleotides derived therefrom, present in a sample. Any method for determining RNA levels can be used. For example, RNA is isolated from a sample and separated on an agarose gel. The separated RNA is then transferred to a solid support, such as a filter. Nucleic acid probes representing one or more markers are then hybridized to the filter by northern hybridization, and the amount of marker-derived RNA is determined. Such determination can be visual, or machine-aided, for example, by use of a densitometer. Another method of determining RNA levels is by use of a dot-blot or a slot-blot. In this method, RNA, or nucleic acid derived therefrom, from a sample is labeled. The RNA or nucleic acid derived therefrom is then hybridized to

a filter containing oligonucleotides derived from one or more marker genes, wherein the oligonucleotides are placed upon the filter at discrete, easily-identifiable locations. Hybridization, or lack thereof, of the labeled RNA to the filter-bound oligonucleotides is determined visually or by densitometer. Polynucleotides can be labeled using a radiolabel
5 or a fluorescent (*i.e.*, visible) label.

These examples are not intended to be limiting; other methods of determining RNA abundance are known in the art.

The level of expression of particular marker genes may also be assessed by determining the level of the specific protein expressed from the marker genes. This can be
10 accomplished, for example, by separation of proteins from a sample on a polyacrylamide gel, followed by identification of specific marker-derived proteins using antibodies in a western blot. Alternatively, proteins can be separated by two-dimensional gel electrophoresis systems. Two-dimensional gel electrophoresis is well-known in the art and typically involves isoelectric focusing along a first dimension followed by SDS-PAGE
15 electrophoresis along a second dimension. *See, e.g.*, Hames et al, 1990, GEL ELECTROPHORESIS OF PROTEINS: A PRACTICAL APPROACH, IRL Press, New York; Shevchenko *et al.*, *Proc. Nat'l Acad. Sci. USA* 93:1440-1445 (1996); Sagliocco *et al.*, *Yeast* 12:1519-1533 (1996); Lander, *Science* 274:536-539 (1996). The resulting electropherograms can be analyzed by numerous techniques, including mass spectrometric
20 techniques, western blotting and immunoblot analysis using polyclonal and monoclonal antibodies.

Alternatively, marker-derived protein levels can be determined by constructing an antibody microarray in which binding sites comprise immobilized, preferably monoclonal, antibodies specific to a plurality of protein species encoded by the
25 cell genome. Preferably, antibodies are present for a substantial fraction of the marker-derived proteins of interest. Methods for making monoclonal antibodies are well known (*see, e.g.*, Harlow and Lane, 1988, ANTIBODIES: A LABORATORY MANUAL, Cold Spring Harbor, New York, which is incorporated in its entirety for all purposes). In one embodiment, monoclonal antibodies are raised against synthetic peptide fragments designed
30 based on genomic sequence of the cell. With such an antibody array, proteins from the cell are contacted to the array, and their binding is assayed with assays known in the art. Generally, the expression, and the level of expression, of proteins of diagnostic or prognostic interest can be detected through immunohistochemical staining of tissue slices or sections.

35

Finally, expression of marker genes in a number of tissue specimens may be characterized using a "tissue array" (Kononen *et al.*, *Nat. Med* 4(7):844-7 (1998)). In a tissue array, multiple tissue samples are assessed on the same microarray. The arrays allow *in situ* detection of RNA and protein levels; consecutive sections allow the analysis of multiple samples simultaneously.

5.5.2 MICROARRAYS

In preferred embodiments, polynucleotide microarrays are used to measure expression so that the expression status of each of the markers above is assessed simultaneously. In a specific embodiment, the invention provides for oligonucleotide or cDNA arrays comprising probes hybridizable to the genes corresponding to each of the marker sets described above (*i.e.*, markers to determine the molecular type or subtype of a tumor; markers to distinguish ER status; markers to distinguish *BRCA1* from sporadic tumors; markers to distinguish patients with good versus patients with poor prognosis; markers to distinguish both ER(+) from ER(-), and *BRCA1* tumors from sporadic tumors; markers to distinguish ER(+) from ER(-), and patients with good prognosis from patients with poor prognosis; markers to distinguish *BRCA1* tumors from sporadic tumors, and patients with good prognosis from patients with poor prognosis; and markers able to distinguish ER(+) from ER(-), *BRCA1* tumors from sporadic tumors, and patients with good prognosis from patients with poor prognosis; and markers unique to each status).

The microarrays provided by the present invention may comprise probes hybridizable to the genes corresponding to markers able to distinguish the status of one, two, or all three of the clinical conditions noted above. In particular, the invention provides polynucleotide arrays comprising probes to a subset or subsets of at least 50, 100, 200, 300, 400, 500, 750, 1,000, 1,250, 1,500, 1,750, 2,000 or 2,250 genetic markers, up to the full set of 2,460 markers, which distinguish ER(+) and ER(-) patients or tumors. The invention also provides probes to subsets of at least 20, 30, 40, 50, 75, 100, 150, 200, 250, 300, 350 or 400 markers, up to the full set of 430 markers, which distinguish between tumors containing a *BRCA1* mutation and sporadic tumors within an ER(-) group of tumors. The invention also provides probes to subsets of at least 20, 30, 40, 50, 75, 100, 150 or 200 markers, up to the full set of 231 markers, which distinguish between patients with good and poor prognosis within sporadic tumors. In a specific embodiment, the array comprises probes to marker sets or subsets directed to any two of the clinical conditions. In a more specific embodiment, the array comprises probes to marker sets or subsets directed to all three clinical conditions.

In yet another specific embodiment, microarrays that are used in the methods disclosed herein optionally comprise markers additional to at least some of the markers listed in Tables 1-6. For example, in a specific embodiment, the microarray is a screening or scanning array as described in Altschuler *et al.*, International Publication WO 02/18646, published March 7, 2002 and Scherer *et al.*, International Publication WO 02/16650, published February 28, 2002. The scanning and screening arrays comprise regularly-spaced, positionally-addressable probes derived from genomic nucleic acid sequence, both expressed and unexpressed. Such arrays may comprise probes corresponding to a subset of, or all of, the markers listed in Tables 1-6, or a subset thereof as described above, and can be used to monitor marker expression in the same way as a microarray containing only markers listed in Tables 1-6.

In yet another specific embodiment, the microarray is a commercially-available cDNA microarray that comprises at least five of the markers listed in Tables 1-6. Preferably, a commercially-available cDNA microarray comprises all of the markers listed in Tables 1-6. However, such a microarray may comprise 5, 10, 15, 25, 50, 100, 150, 250, 500, 1000 or more of the markers in any of Tables 1-6, up to the maximum number of markers in a Table, and may comprise all of the markers in any one of Tables 1-6 and a subset of another of Tables 1-6, or subsets of each as described above. In a specific embodiment of the microarrays used in the methods disclosed herein, the markers that are all or a portion of Tables 1-6 make up at least 50%, 60%, 70%, 80%, 90%, 95% or 98% of the probes on the microarray.

General methods pertaining to the construction of microarrays comprising the marker sets and/or subsets above are described in the following sections.

5.5.2.1 CONSTRUCTION OF MICROARRAYS

Microarrays are prepared by selecting probes which comprise a polynucleotide sequence, and then immobilizing such probes to a solid support or surface. For example, the probes may comprise DNA sequences, RNA sequences, or copolymer sequences of DNA and RNA. The polynucleotide sequences of the probes may also comprise DNA and/or RNA analogues, or combinations thereof. For example, the polynucleotide sequences of the probes may be full or partial fragments of genomic DNA. The polynucleotide sequences of the probes may also be synthesized nucleotide sequences, such as synthetic oligonucleotide sequences. The probe sequences can be synthesized either enzymatically *in vivo*, enzymatically *in vitro* (e.g., by PCR), or non-enzymatically *in vitro*.

The probe or probes used in the methods of the invention are preferably immobilized to a solid support which may be either porous or non-porous. For example, the probes of the invention may be polynucleotide sequences which are attached to a nitrocellulose or nylon membrane or filter covalently at either the 3' or the 5' end of the polynucleotide. Such hybridization probes are well known in the art (see, *e.g.*, Sambrook et al., *MOLECULAR CLONING - A LABORATORY MANUAL* (2ND ED.), Vols. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York (1989). Alternatively, the solid support or surface may be a glass or plastic surface. In a particularly preferred embodiment, hybridization levels are measured to microarrays of probes consisting of a solid phase on the surface of which are immobilized a population of polynucleotides, such as a population of DNA or DNA mimics, or, alternatively, a population of RNA or RNA mimics. The solid phase may be a nonporous or, optionally, a porous material such as a gel.

In preferred embodiments, a microarray comprises a support or surface with an ordered array of binding (*e.g.*, hybridization) sites or "probes" each representing one of the markers described herein. Preferably the microarrays are addressable arrays, and more preferably positionally addressable arrays. More specifically, each probe of the array is preferably located at a known, predetermined position on the solid support such that the identity (*i.e.*, the sequence) of each probe can be determined from its position in the array (*i.e.*, on the support or surface). In preferred embodiments, each probe is covalently attached to the solid support at a single site.

Microarrays can be made in a number of ways, of which several are described below. However produced, microarrays share certain characteristics. The arrays are reproducible, allowing multiple copies of a given array to be produced and easily compared with each other. Preferably, microarrays are made from materials that are stable under binding (*e.g.*, nucleic acid hybridization) conditions. The microarrays are preferably small, *e.g.*, between 1 cm² and 25 cm², between 12 cm² and 13 cm², or 3 cm². However, larger arrays are also contemplated and may be preferable, *e.g.*, for use in screening arrays. Preferably, a given binding site or unique set of binding sites in the microarray will specifically bind (*e.g.*, hybridize) to the product of a single gene in a cell (*e.g.*, to a specific mRNA, or to a specific cDNA derived therefrom). However, in general, other related or similar sequences will cross hybridize to a given binding site.

The microarrays of the present invention include one or more test probes, each of which has a polynucleotide sequence that is complementary to a subsequence of RNA or DNA to be detected. Preferably, the position of each probe on the solid surface is known. Indeed, the microarrays are preferably positionally addressable arrays. Specifically, each

probe of the array is preferably located at a known, predetermined position on the solid support such that the identity (*i.e.*, the sequence) of each probe can be determined from its position on the array (*i.e.*, on the support or surface).

According to the invention, the microarray is an array (*i.e.*, a matrix) in which each position represents one of the markers described herein. For example, each position can contain a DNA or DNA analogue based on genomic DNA to which a particular RNA or cDNA transcribed from that genetic marker can specifically hybridize. The DNA or DNA analogue can be, *e.g.*, a synthetic oligomer or a gene fragment. In one embodiment, probes representing each of the markers is present on the array. In a preferred embodiment, the array comprises the 550 of the 2,460 RE-status markers, 70 of the *BRCA1*/sporadic markers, and all 231 of the prognosis markers.

5.5.2.2 PREPARING PROBES FOR MICROARRAYS

As noted above, the “probe” to which a particular polynucleotide molecule specifically hybridizes according to the invention contains a complementary genomic polynucleotide sequence. The probes of the microarray preferably consist of nucleotide sequences of no more than 1,000 nucleotides. In some embodiments, the probes of the array consist of nucleotide sequences of 10 to 1,000 nucleotides. In a preferred embodiment, the nucleotide sequences of the probes are in the range of 10-200 nucleotides in length and are genomic sequences of a species of organism, such that a plurality of different probes is present, with sequences complementary and thus capable of hybridizing to the genome of such a species of organism, sequentially tiled across all or a portion of such genome. In other specific embodiments, the probes are in the range of 10-30 nucleotides in length, in the range of 10-40 nucleotides in length, in the range of 20-50 nucleotides in length, in the range of 40-80 nucleotides in length, in the range of 50-150 nucleotides in length, in the range of 80-120 nucleotides in length, and most preferably are 60 nucleotides in length.

The probes may comprise DNA or DNA “mimics” (*e.g.*, derivatives and analogues) corresponding to a portion of an organism’s genome. In another embodiment, the probes of the microarray are complementary RNA or RNA mimics. DNA mimics are polymers composed of subunits capable of specific, Watson-Crick-like hybridization with DNA, or of specific hybridization with RNA. The nucleic acids can be modified at the base moiety, at the sugar moiety, or at the phosphate backbone. Exemplary DNA mimics include, *e.g.*, phosphorothioates.

DNA can be obtained, *e.g.*, by polymerase chain reaction (PCR) amplification of genomic DNA or cloned sequences. PCR primers are preferably chosen based on a known

sequence of the genome that will result in amplification of specific fragments of genomic DNA. Computer programs that are well known in the art are useful in the design of primers with the required specificity and optimal amplification properties, such as *Oligo* version 5.0 (National Biosciences). Typically each probe on the microarray will be between 10 bases
5 and 50,000 bases, usually between 300 bases and 1,000 bases in length. PCR methods are well known in the art, and are described, for example, in Innis *et al.*, eds., PCR PROTOCOLS: A GUIDE TO METHODS AND APPLICATIONS, Academic Press Inc., San Diego, CA (1990). It will be apparent to one skilled in the art that controlled robotic systems are useful for isolating and amplifying nucleic acids.

- 10 An alternative, preferred means for generating the polynucleotide probes of the microarray is by synthesis of synthetic polynucleotides or oligonucleotides, *e.g.*, using N-phosphonate or phosphoramidite chemistries (Froehler *et al.*, *Nucleic Acid Res.* 14:5399-5407 (1986); McBride *et al.*, *Tetrahedron Lett.* 24:246-248 (1983)). Synthetic sequences are typically between about 10 and about 500 bases in length, more typically between about
15 20 and about 100 bases, and most preferably between about 40 and about 70 bases in length. In some embodiments, synthetic nucleic acids include non-natural bases, such as, but by no means limited to, inosine. As noted above, nucleic acid analogues may be used as binding sites for hybridization. An example of a suitable nucleic acid analogue is peptide nucleic acid (*see, e.g.*, Egholm *et al.*, *Nature* 363:566-568 (1993); U.S. Patent No. 5,539,083).
- 20 Probes are preferably selected using an algorithm that takes into account binding energies, base composition, sequence complexity, cross-hybridization binding energies, and secondary structure (*see* Friend *et al.*, International Patent Publication WO 01/05935, published January 25, 2001; Hughes *et al.*, *Nat. Biotech.* 19:342-7 (2001)).

- A skilled artisan will also appreciate that positive control probes, *e.g.*, probes known
25 to be complementary and hybridizable to sequences in the target polynucleotide molecules, and negative control probes, *e.g.*, probes known to not be complementary and hybridizable to sequences in the target polynucleotide molecules, should be included on the array. In one embodiment, positive controls are synthesized along the perimeter of the array. In another embodiment, positive controls are synthesized in diagonal stripes across the array. In still
30 another embodiment, the reverse complement for each probe is synthesized next to the position of the probe to serve as a negative control. In yet another embodiment, sequences from other species of organism are used as negative controls or as "spike-in" controls.

5.5.2.3 ATTACHING PROBES TO THE SOLID SURFACE

The probes are attached to a solid support or surface, which may be made, *e.g.*, from glass, plastic (*e.g.*, polypropylene, nylon), polyacrylamide, nitrocellulose, gel, or other porous or nonporous material. A preferred method for attaching the nucleic acids to a surface is by printing on glass plates, as is described generally by Schena *et al.*, *Science* 270:467-470 (1995). This method is especially useful for preparing microarrays of cDNA (See also, DeRisi *et al.*, *Nature Genetics* 14:457-460 (1996); Shalon *et al.*, *Genome Res.* 6:639-645 (1996); and Schena *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 93:10539-11286 (1995)).

A second preferred method for making microarrays is by making high-density oligonucleotide arrays. Techniques are known for producing arrays containing thousands of oligonucleotides complementary to defined sequences, at defined locations on a surface using photolithographic techniques for synthesis *in situ* (see, Fodor *et al.*, 1991, *Science* 251:767-773; Pease *et al.*, 1994, *Proc. Natl. Acad. Sci. U.S.A.* 91:5022-5026; Lockhart *et al.*, 1996, *Nature Biotechnology* 14:1675; U.S. Patent Nos. 5,578,832; 5,556,752; and 5,510,270) or other methods for rapid synthesis and deposition of defined oligonucleotides (Blanchard *et al.*, *Biosensors & Bioelectronics* 11:687-690). When these methods are used, oligonucleotides (*e.g.*, 60-mers) of known sequence are synthesized directly on a surface such as a derivatized glass slide. Usually, the array produced is redundant, with several oligonucleotide molecules per RNA.

Other methods for making microarrays, *e.g.*, by masking (Maskos and Southern, 1992, *Nuc. Acids. Res.* 20:1679-1684), may also be used. In principle, and as noted *supra*, any type of array, for example, dot blots on a nylon hybridization membrane (see Sambrook *et al.*, *MOLECULAR CLONING - A LABORATORY MANUAL* (2ND ED.), Vols. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York (1989)) could be used. However, as will be recognized by those skilled in the art, very small arrays will frequently be preferred because hybridization volumes will be smaller.

In one embodiment, the arrays of the present invention are prepared by synthesizing polynucleotide probes on a support. In such an embodiment, polynucleotide probes are attached to the support covalently at either the 3' or the 5' end of the polynucleotide.

In a particularly preferred embodiment, microarrays of the invention are manufactured by means of an ink jet printing device for oligonucleotide synthesis, *e.g.*, using the methods and systems described by Blanchard in U.S. Pat. No. 6,028,189; Blanchard *et al.*, 1996, *Biosensors and Bioelectronics* 11:687-690; Blanchard, 1998, in *SYNTHETIC DNA ARRAYS IN GENETIC ENGINEERING*, Vol. 20, J.K. Setlow, Ed., Plenum Press, New York at pages 111-123. Specifically, the oligonucleotide probes in such

microarrays are preferably synthesized in arrays, *e.g.*, on a glass slide, by serially depositing individual nucleotide bases in "microdroplets" of a high surface tension solvent such as propylene carbonate. The microdroplets have small volumes (*e.g.*, 100 pL or less, more preferably 50 pL or less) and are separated from each other on the microarray (*e.g.*, by hydrophobic domains) to form circular surface tension wells which define the locations of the array elements (*i.e.*, the different probes). Microarrays manufactured by this ink-jet method are typically of high density, preferably having a density of at least about 2,500 different probes per 1 cm². The polynucleotide probes are attached to the support covalently at either the 3' or the 5' end of the polynucleotide.

10

5.5.2.4 TARGET POLYNUCLEOTIDE MOLECULES

The polynucleotide molecules which may be analyzed by the present invention (the "target polynucleotide molecules") may be from any clinically relevant source, but are expressed RNA or a nucleic acid derived therefrom (*e.g.*, cDNA or amplified RNA derived from cDNA that incorporates an RNA polymerase promoter), including naturally occurring nucleic acid molecules, as well as synthetic nucleic acid molecules. In one embodiment, the target polynucleotide molecules comprise RNA, including, but by no means limited to, total cellular RNA, poly(A)⁺ messenger RNA (mRNA) or fraction thereof, cytoplasmic mRNA, or RNA transcribed from cDNA (*i.e.*, cRNA; see, *e.g.*, Linsley & Schelter, U.S. Patent Application No. 09/411,074, filed October 4, 1999, or U.S. Patent Nos. 5,545,522, 5,891,636, or 5,716,785). Methods for preparing total and poly(A)⁺ RNA are well known in the art, and are described generally, *e.g.*, in Sambrook *et al.*, MOLECULAR CLONING - A LABORATORY MANUAL (2ND ED.), Vols. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York (1989). In one embodiment, RNA is extracted from cells of the various types of interest in this invention using guanidinium thiocyanate lysis followed by CsCl centrifugation (Chirgwin *et al.*, 1979, *Biochemistry* 18:5294-5299). In another embodiment, total RNA is extracted using a silica gel-based column, commercially available examples of which include RNeasy (Qiagen, Valencia, California) and StrataPrep (Stratagene, La Jolla, California). In an alternative embodiment, which is preferred for *S. cerevisiae*, RNA is extracted from cells using phenol and chloroform, as described in Ausubel *et al.*, eds., 1989, CURRENT PROTOCOLS IN MOLECULAR BIOLOGY, Vol III, Green Publishing Associates, Inc., John Wiley & Sons, Inc., New York, at pp. 13.12.1-13.12.5). Poly(A)⁺ RNA can be selected, *e.g.*, by selection with oligo-dT cellulose or, alternatively, by oligo-dT primed reverse transcription of total cellular RNA. In one embodiment, RNA can be fragmented by methods known in the art, *e.g.*, by incubation with ZnCl₂, to generate

35

fragments of RNA. In another embodiment, the polynucleotide molecules analyzed by the invention comprise cDNA, or PCR products of amplified RNA or cDNA.

In one embodiment, total RNA, mRNA, or nucleic acids derived therefrom, is isolated from a sample taken from a person afflicted with breast cancer. Target
5 polynucleotide molecules that are poorly expressed in particular cells may be enriched using normalization techniques (Bonaldo *et al.*, 1996, *Genome Res.* 6:791-806).

As described above, the target polynucleotides are detectably labeled at one or more nucleotides. Any method known in the art may be used to detectably label the target polynucleotides. Preferably, this labeling incorporates the label uniformly along the length
10 of the RNA, and more preferably, the labeling is carried out at a high degree of efficiency. One embodiment for this labeling uses oligo-dT primed reverse transcription to incorporate the label; however, conventional methods of this method are biased toward generating 3' end fragments. Thus, in a preferred embodiment, random primers (*e.g.*, 9-mers) are used in reverse transcription to uniformly incorporate labeled nucleotides over the full length of the
15 target polynucleotides. Alternatively, random primers may be used in conjunction with PCR methods or T7 promoter-based *in vitro* transcription methods in order to amplify the target polynucleotides.

In a preferred embodiment, the detectable label is a luminescent label. For example, fluorescent labels, bio-luminescent labels, chemi-luminescent labels, and colorimetric labels
20 may be used in the present invention. In a highly preferred embodiment, the label is a fluorescent label, such as a fluorescein, a phosphor, a rhodamine, or a polymethine dye derivative. Examples of commercially available fluorescent labels include, for example, fluorescent phosphoramidites such as FluorePrime (Amersham Pharmacia, Piscataway, N.J.), Fluoredate (Millipore, Bedford, Mass.), FAM (ABI, Foster City, Calif.), and Cy3 or
25 Cy5 (Amersham Pharmacia, Piscataway, N.J.). In another embodiment, the detectable label is a radiolabeled nucleotide.

In a further preferred embodiment, target polynucleotide molecules from a patient sample are labeled differentially from target polynucleotide molecules of a standard. The standard can comprise target polynucleotide molecules from normal individuals (*i.e.*, those
30 not afflicted with breast cancer). In a highly preferred embodiment, the standard comprises target polynucleotide molecules pooled from samples from normal individuals or tumor samples from individuals having sporadic-type breast tumors. In another embodiment, the target polynucleotide molecules are derived from the same individual, but are taken at different time points, and thus indicate the efficacy of a treatment by a change in expression
35 of the markers, or lack thereof, during and after the course of treatment (*i.e.*, chemotherapy,

radiation therapy or cryotherapy), wherein a change in the expression of the markers from a poor prognosis pattern to a good prognosis pattern indicates that the treatment is efficacious. In this embodiment, different timepoints are differentially labeled.

5

5.5.2.5 HYBRIDIZATION TO MICROARRAYS

Nucleic acid hybridization and wash conditions are chosen so that the target polynucleotide molecules specifically bind or specifically hybridize to the complementary polynucleotide sequences of the array, preferably to a specific array site, wherein its complementary DNA is located.

10

Arrays containing double-stranded probe DNA situated thereon are preferably subjected to denaturing conditions to render the DNA single-stranded prior to contacting with the target polynucleotide molecules. Arrays containing single-stranded probe DNA (*e.g.*, synthetic oligodeoxyribonucleic acids) may need to be denatured prior to contacting with the target polynucleotide molecules, *e.g.*, to remove hairpins or dimers which form due to self complementary sequences.

15

Optimal hybridization conditions will depend on the length (*e.g.*, oligomer versus polynucleotide greater than 200 bases) and type (*e.g.*, RNA, or DNA) of probe and target nucleic acids. One of skill in the art will appreciate that as the oligonucleotides become shorter, it may become necessary to adjust their length to achieve a relatively uniform melting temperature for satisfactory hybridization results. General parameters for specific (*i.e.*, stringent) hybridization conditions for nucleic acids are described in Sambrook *et al.*, MOLECULAR CLONING - A LABORATORY MANUAL (2ND ED.), Vols. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York (1989), and in Ausubel *et al.*, CURRENT PROTOCOLS IN MOLECULAR BIOLOGY, vol. 2, Current Protocols Publishing, New York (1994). Typical hybridization conditions for the cDNA microarrays of Schena *et al.* are hybridization in 5 X SSC plus 0.2% SDS at 65 °C for four hours, followed by washes at 25 °C in low stringency wash buffer (1 X SSC plus 0.2% SDS), followed by 10 minutes at 25 °C in higher stringency wash buffer (0.1 X SSC plus 0.2% SDS) (Schena *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 93:10614 (1993)). Useful hybridization conditions are also provided in, *e.g.*, Tijessen, 1993, HYBRIDIZATION WITH NUCLEIC ACID PROBES, Elsevier Science Publishers B.V.; and Kricka, 1992, NONISOTOPIC DNA PROBE TECHNIQUES, Academic Press, San Diego, CA.

20

25

30

35

Particularly preferred hybridization conditions include hybridization at a temperature at or near the mean melting temperature of the probes (*e.g.*, within 5 °C, more preferably

within 2 °C) in 1 M NaCl, 50 mM MES buffer (pH 6.5), 0.5% sodium sarcosine and 30% formamide.

5.5.2.6 SIGNAL DETECTION AND DATA ANALYSIS

5 When fluorescently labeled probes are used, the fluorescence emissions at each site of a microarray may be, preferably, detected by scanning confocal laser microscopy. In one embodiment, a separate scan, using the appropriate excitation line, is carried out for each of the two fluorophores used. Alternatively, a laser may be used that allows simultaneous specimen illumination at wavelengths specific to the two fluorophores and emissions from
10 the two fluorophores can be analyzed simultaneously (*see Shalon et al.*, 1996, "A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization," *Genome Research* 6:639-645, which is incorporated by reference in its entirety for all purposes). In a preferred embodiment, the arrays are scanned with a laser fluorescent scanner with a computer controlled X-Y stage and a microscope objective.
15 Sequential excitation of the two fluorophores is achieved with a multi-line, mixed gas laser and the emitted light is split by wavelength and detected with two photomultiplier tubes. Fluorescence laser scanning devices are described in Schena *et al.*, *Genome Res.* 6:639-645 (1996), and in other references cited herein. Alternatively, the fiber-optic bundle described by Ferguson *et al.*, *Nature Biotech.* 14:1681-1684 (1996), may be used to monitor mRNA
20 abundance levels at a large number of sites simultaneously.

 Signals are recorded and, in a preferred embodiment, analyzed by computer, *e.g.*, using a 12 or 16 bit analog to digital board. In one embodiment the scanned image is despeckled using a graphics program (*e.g.*, Hijaak Graphics Suite) and then analyzed using an image gridding program that creates a spreadsheet of the average hybridization at each
25 wavelength at each site. If necessary, an experimentally determined correction for "cross talk" (or overlap) between the channels for the two fluors may be made. For any particular hybridization site on the transcript array, a ratio of the emission of the two fluorophores can be calculated. The ratio is independent of the absolute expression level of the cognate gene, but is useful for genes whose expression is significantly modulated in association with the
30 different breast cancer-related condition.

5.6 COMPUTER-FACILITATED ANALYSIS

 The present invention further provides for kits comprising the marker sets above. In a preferred embodiment, the kit contains a microarray ready for hybridization to
35 target polynucleotide molecules, plus software for the data analyses described above..

The analytic methods described in the previous sections can be implemented by use of the following computer systems and according to the following programs and methods. A Computer system comprises internal components linked to external components. The internal components of a typical computer system include a processor
5 element interconnected with a main memory. For example, the computer system can be an Intel 8086-, 80386-, 80486-, Pentium™, or Pentium™-based processor with preferably 32 MB or more of main memory.

The external components may include mass storage. This mass storage can be one or more hard disks (which are typically packaged together with the processor and
10 memory). Such hard disks are preferably of 1 GB or greater storage capacity. Other external components include a user interface device, which can be a monitor, together with an inputting device, which can be a "mouse", or other graphic input devices, and/or a keyboard. A printing device can also be attached to the computer.

Typically, a computer system is also linked to network link, which can be
15 part of an Ethernet link to other local computer systems, remote computer systems, or wide area communication networks, such as the Internet. This network link allows the computer system to share data and processing tasks with other computer systems.

Loaded into memory during operation of this system are several software components, which are both standard in the art and special to the instant invention. These
20 software components collectively cause the computer system to function according to the methods of this invention. These software components are typically stored on the mass storage device. A software component comprises the operating system, which is responsible for managing computer system and its network interconnections. This operating system can be, for example, of the Microsoft Windows® family, such as
25 Windows 3.1, Windows 95, Windows 98, Windows 2000, or Windows NT. The software component represents common languages and functions conveniently present on this system to assist programs implementing the methods specific to this invention. Many high or low level computer languages can be used to program the analytic methods of this invention. Instructions can be interpreted during run-time or compiled. Preferred languages include C/
30 C++, FORTRAN and JAVA. Most preferably, the methods of this invention are programmed in mathematical software packages that allow symbolic entry of equations and high-level specification of processing, including some or all of the algorithms to be used, thereby freeing a user of the need to procedurally program individual equations or algorithms. Such packages include Mathlab from Mathworks (Natick, MA), Mathematica®
35 from Wolfram Research (Champaign, IL), or S-Plus® from Math Soft (Cambridge, MA).

Specifically, the software component includes the analytic methods of the invention as programmed in a procedural language or symbolic package.

The software to be included with the kit comprises the data analysis methods of the invention as disclosed herein. In particular, the software may include mathematical
5 routines for marker discovery, including the calculation of correlation coefficients between clinical categories (*i.e.*, ER status) and marker expression. The software may also include mathematical routines for calculating the correlation between sample marker expression and control marker expression, using array-generated fluorescence data, to determine the clinical classification of a sample.

10 In an exemplary implementation, to practice the methods of the present invention, a user first loads experimental data into the computer system. These data can be directly entered by the user from a monitor, keyboard, or from other computer systems linked by a network connection, or on removable storage media such as a CD-ROM, floppy disk (not illustrated), tape drive (not illustrated), ZIP® drive (not illustrated) or through the
15 network. Next the user causes execution of expression profile analysis software which performs the methods of the present invention.

In another exemplary implementation, a user first loads experimental data and/or databases into the computer system. This data is loaded into the memory from the storage media or from a remote computer, preferably from a dynamic geneset database
20 system, through the network. Next the user causes execution of software that performs the steps of the present invention.

Alternative computer systems and software for implementing the analytic methods of this invention will be apparent to one of skill in the art and are intended to be comprehended within the accompanying claims. In particular, the accompanying claims are
25 intended to include the alternative program structures for implementing the methods of this invention that will be readily apparent to one of skill in the art.

6. EXAMPLES

Materials And Methods

30 117 tumor samples from breast cancer patients were collected. RNA samples were then prepared, and each RNA sample was profiled using inkjet-printed microarrays. Marker genes were then identified based on expression patterns; these genes were then used to train classifiers, which used these marker genes to classify tumors into diagnostic and prognostic categories. Finally, these marker genes were used to predict the
35 diagnostic and prognostic outcome for a group of individuals..

1. Sample collection

117 breast cancer patients treated at The Netherlands Cancer Institute / Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands, were selected on the basis of the following clinical criteria (data extracted from the medical records of the NKI/AvL Tumor Register, Biometrics Department).

Group 1 (n=97, 78 for training, 19 for independent tests) was selected on the basis of: (1) primary invasive breast carcinoma <5 cm (T1 or T2); (2) no axillary metastases (N0); (3) age at diagnosis <55 years; (4) calendar year of diagnosis 1983-1996; and (5) no prior malignancies (excluding carcinoma in situ of the cervix or basal cell carcinoma of the skin). All patients were treated by modified radical mastectomy (n=34) or breast conserving treatment (n=64), including axillary lymph node dissection. Breast conserving treatment consisted of excision of the tumor, followed by radiation of the whole breast to a dosis of 50 Gy, followed by a boost varying from 15 to 25 Gy. Five patients received adjuvant systemic therapy consisting of chemotherapy (n=3) or hormonal therapy (n=2), all other patients did not receive additional treatment. All patients were followed at least annually for a period of at least 5 years. Patient follow-up information was extracted from the Tumor Registry of the Biometrics Department.

Group 2 (n=20) was selected as: (1) carriers of a germline mutation in *BRCA1* or *BRCA2*; and (2) having primary invasive breast carcinoma. No selection or exclusion was made based on tumor size, lymph node status, age at diagnosis, calendar year of diagnosis, other malignancies. Germline mutation status was known prior to this research protocol.

Information about individual from which tumor samples were collected include: year of birth; sex; whether the individual is pre- or post-menopausal; the year of diagnosis; the number of positive lymph nodes and the total number of nodes; whether there was surgery, and if so, whether the surgery was breast-conserving or radical; whether there was radiotherapy, chemotherapy or hormonal therapy. The tumor was graded according to the formula $P=TNM$, where T is the tumor size (on a scale of 0-5); N is the number of nodes that are positive (on a scale of 0-4); and M is metastases (0 = absent, 1 = present). The tumor was also classified according to stage, tumor type (*in situ* or invasive; lobular or ductal; grade) and the presence or absence of the estrogen and progesterone receptors. The progression of the cancer was described by (where applicable): distant metastases; year of distant metastases, year of death, year of last follow-up; and *BRCA1* genotype.

2. Tumors:

Germline mutation testing of *BRCA1* and *BRCA2* on DNA isolated from peripheral blood lymphocytes includes mutation screening by a Protein Truncation Test (PTT) of exon 11 of *BRCA1* and exon 10 and 11 of *BRCA2*, deletion PCR of *BRCA1*
5 genomic deletion of exon 13 and 22, as well Denaturing Gradient Gel Electrophoresis (DGGE) of the remaining exons. Aberrant bands were all confirmed by genomic sequencing analyzed on a ABI3700 automatic sequencer and confirmed on a independent DNA sample.

From all, tumor material was snap frozen in liquid nitrogen within one hour after surgery.
10 Of the frozen tumor material an H&E (hematoxylin-eosin) stained section was prepared prior to and after cutting slides for RNA isolation. These H&E frozen sections were assessed for the percentage of tumor cells; only samples with >50% tumor cells were selected for further study.

For all tumors, surgical specimens fixed in formaldehyde and embedded in
15 paraffin were evaluated according to standard histopathological procedures. H&E stained paraffin sections were examined to assess tumor type (e.g., ductal or lobular according to the WHO classification); to assess histologic grade according the method described by Elston and Ellis (grade 1-3); and to assess the presence of lymphangio-invasive growth and the presence of an extensive lymphocytic infiltrate. All histologic factors were
20 independently assessed by two pathologists (MV and JL); consensus on differences was reached by examining the slides together. A representative slide of each tumor was used for immunohistochemical staining with antibodies directed against the estrogen- and progesterone receptor by standard procedures. The staining result was scored as the percentage of positively staining nuclei (0%, 10%, 20%, etc., up to 100%).

25

3. Amplification, labeling, and hybridization

The outline for the production of marker-derived nucleic acids and hybridization of the nucleic acids to a microarray are outlined in FIG. 2. 30 frozen sections of 30 μ M thickness were used for total RNA isolation of each snap frozen tumor specimen.
30 Total RNA was isolated with RNeasyTM B (Qiagen Scientific, Venendaal, The Netherlands) according to the manufacturers protocol, including homogenization of the tissue using a Polytron PT-MR2100 (Merck, Amsterdam, The Netherlands) and finally dissolved in RNase-free H₂O. The quality of the total RNA was assessed by A260/A280 ratio and had to be between 1.7 and 2.1 as well as visual inspection of the RNA on an
35 agarose gel which should indicate a stronger 28S ribosomal RNA band compared to the 18S

ribosomal RNA band. subsequently, 25µg of total RNA was DNase treated using the Qiagen RNase-free DNase kit and RNeasy spin columns (Qiagen Inc, GmbH, Germany) according to the manufacturers protocol. DNase treated total RNA was dissolved in RNase-free H₂O to a final concentration of 0.2µg/µl.

- 5 5µg total RNA was used as input for cRNA synthesis. An oligo-dT primer containing a T7 RNA polymerase promoter sequence was used to prime first strand cDNA synthesis, and random primers (pdN6) were used to prime second strand cDNA synthesis by MMLV reverse transcriptase. This reaction yielded a double-stranded cDNA that contained the T7 RNA polymerase (T7RNAP) promoter. The double-stranded cDNA was then
10 transcribed into cRNA by T7RNAP.

- cRNA was labeled with Cy3 or Cy5 dyes using a two-step process. First, allylamine-derivitized nucleotides were enzymatically incorporated into cRNA products. For cRNA labeling, a 3:1 mixture of 5-(3-Aminoallyl)uridine 5'-triphosphate (Sigma) and UTP was substituted for UTP in the in vitro transcription (IVT) reaction. Allylamine-
15 derivitized cRNA products were then reacted with N-hydroxy succinimide esters of Cy3 or Cy5 (CyDye, Amersham Pharmacia Biotech). 5µg Cy5-labeled cRNA from one breast cancer patient was mixed with the same amount of Cy3-labeled product from a pool of equal amount of cRNA from each individual sporadic patient.

- Microarray hybridizations were done in duplicate with fluor reversals.
20 Before hybridization, labeled cRNAs were fragmented to an average size of ~50-100nt by heating at 60 °C in the presence of 10 mM ZnCl₂. Fragmented cRNAs were added to hybridization buffer containing 1 M NaCl, 0.5% sodium sarcosine and 50 mM MES, pH 6.5, which stringency was regulated by the addition of formamide to a final concentration of 30%. Hybridizations were carried out in a final volume of 3 mls at 40 °C on a rotating
25 platform in a hybridization oven (Robbins Scientific) for 48h. After hybridization, slides were washed and scanned using a confocal laser scanner (Agilent Technologies). Fluorescence intensities on scanned images were quantified, normalized and corrected.

4. Pooling of samples

- 30 The reference cRNA pool was formed by pooling equal amount of cRNAs from each individual sporadic patient, for a total of 78 tumors.

5. 25k human microarray

- Surface-bound oligonucleotides were synthesized essentially as proposed by
35 Blanchard *et al.*, *Biosens. Bioelectron.* 6(7):687-690 (1996); see also Hughes *et al.*, *Nature*

Biotech. 19(4):342-347 (2000). Hydrophobic glass surfaces (3 inches by 3 inches) containing exposed hydroxyl groups were used as substrates for nucleotide synthesis. Phosphoramidite monomers were delivered to computer-defined positions on the glass surfaces using ink-jet printer heads. Unreacted monomers were then washed away and the
5 ends of the extended oligonucleotides were deprotected. This cycle of monomer coupling, washing and deprotection was repeated for each desired layer of nucleotide synthesis. Oligonucleotide sequences to be printed were specified by computer files.

Microarrays containing approximately 25,000 human gene sequences (Hu25K microarrays) were used for this study. Sequences for microarrays were selected
10 from RefSeq (a collection of non-redundant mRNA sequences, located on the Internet at nlm.nih.gov/LocusLink/refseq.html) and Phil Green EST contigs, which is a collection of EST contigs assembled by Dr. Phil Green et al at the University of Washington (Ewing and Green, *Nat. Genet.* 25(2):232-4 (2000)), available on the Internet at phrap.org/est_assembly/index.html. Each mRNA or EST contig was represented on Hu25K microarray by a single
15 60mer oligonucleotide essentially as described in Hughes *et al.*, *Nature Biotech.* 19(4):342-347 and in International Publication WO 01/06013, published January 25, 2001, and in International Publication WO 01/05935, published January 25, 2001, except that the rules for oligo screening were modified to remove oligonucleotides with more than 30%C or with 6 or more contiguous C residues.

20

Example 1: Differentially regulated gene sets and overall expression patterns of breast cancer tumors

Of the approximately 25,000 sequences represented on the microarray, a group of approximately 5,000 genes that were significantly regulated across the group of samples
25 was selected. A gene was determined to be significantly differentially regulated with cancer of the breast if it showed more than two-fold of transcript changes as compared to a sporadic tumor pool, and if the p-value for differential regulation (Hughes *et al.*, *Cell* 102:109-126 (2000)) was less than 0.01 either upwards or downwards in at least five out of 98 tumor samples.

30

An unsupervised clustering algorithm allowed us to cluster patients based on their similarities measured over this set of ~5,000 significant genes. The similarity measure between two patients x and y is defined as

35

$$S = 1 - \left[\frac{\sum_{i=1}^{N_y} \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sigma_{x_i} \sigma_{y_i}}}{\sqrt{\sum_{i=1}^{N_y} \left(\frac{x_i - \bar{x}}{\sigma_{x_i}} \right)^2 \sum_{i=1}^{N_y} \left(\frac{y_i - \bar{y}}{\sigma_{y_i}} \right)^2}} \right] \quad \text{Equation (5)}$$

5

In Equation (5), \mathcal{X} and \mathcal{Y} are two patients with components of log ratio x_i and y_i , $i=1, \dots$,

10 $N=5,100$. Associated with every value x_i is error σ_{x_i} . The smaller the value σ_{x_i} , the more

reliable the measurement x_i . $\bar{x} = \sum_{i=1}^{N_y} \frac{x_i}{\sigma_{x_i}^2} / \sum_{i=1}^{N_y} \frac{1}{\sigma_{x_i}^2}$ is the error-weighted arithmetic mean.

15 The use of correlation as similarity metric emphasizes the importance of co-regulation in clustering rather than the amplitude of regulations.

The set of approximately 5,000 genes can be clustered based on their similarities measured over the group of 98 tumor samples. The similarity measure between two genes was defined in the same way as in Equation (1) except that now for each gene, there are 98 components of log ratio measurements.

20 The result of such a two-dimensional clustering is displayed in FIG 3. Two distinctive patterns emerge from the clustering. The first pattern consists of a group of patients in the lower part of the plot whose regulations are very different from the sporadic pool. The other pattern is made of a group of patients in the upper part of the plot whose expressions are only moderately regulated in comparison with the sporadic pool. These

25 dominant patterns suggest that the tumors can be unambiguously divided into two distinct types based on this set of ~5,000 significant genes.

To help understand these patterns, they were associated with estrogen-receptor (ER), proestrogen receptor (PR), tumor grade, presence of lymphocytic infiltrate, and angiogenesis (FIG. 3). The lower group in FIG 3, which features the dominant pattern,

30 consists of 36 patients. Of the 39 ER-negative patients, 34 patients are clustered together in this group. From FIG. 4, it was observed that the expression of estrogen receptor alpha gene *ESR1* and a large group of co-regulated genes are consistent with this expression pattern.

35 From FIG. 3 and FIG. 4, it was concluded that gene expression patterns can be used to classify tumor samples into subgroups of diagnostic interest. Thus, genes co-

regulated across 98 tumor samples contain information about the molecular basis of breast cancers. The combination of clinical data and microarray measured gene abundance of *ESR1* demonstrates that the distinct types are related to, or at least are reported by, the ER status.

5

Example 2: Identification of Genetic Markers Distinguishing Estrogen Receptor (+) From Estrogen Receptor (-) Patients

The results described in this Example allow the identification of expression marker genes that differentiate two major types of tumor cells: "ER-negative" group and "ER-positive" group. The differentiation of samples by ER(+) status was accomplished in three steps: (1) identification of a set of candidate marker genes that correlate with ER level; (2) rank-ordering these candidate genes by strength of correlation; (3) optimization of the number of marker genes; and (4) classifying samples based on these marker genes.

15

1. Selection of candidate discriminating genes

In the first step, a set of candidate discriminating genes was identified based on gene expression data of training samples. Specifically, we calculated the correlation coefficients ρ between the category numbers or ER level and logarithmic expression ratio \vec{r} across all the samples for each individual gene:

20

$$\rho = (\vec{c} \bullet \vec{r}) / (\|\vec{c}\| \cdot \|\vec{r}\|)$$

Equation (2)

The histogram of resultant correlation coefficients is shown in FIG. 5A as a gray line. While the amplitude of correlation or anti-correlation is small for the majority of genes, the amplitude for some genes is as great as 0.5. Genes whose expression ratios either correlate or anti-correlate well with the diagnostic category of interest are used as reporter genes for the category.

Genes having a correlation coefficient larger than 0.3 ("correlated genes") or less than -0.3 ("anti-correlated genes") were selected as reporter genes. The threshold of 0.3 was selected based on the correlation distribution for cases where there is no real correlation (one can use permutations to determine this distribution). Statistically, this distribution width depends upon the number of samples used in the correlation calculation.

The distribution width for control cases (no real correlation) is approximately $1/\sqrt{n-3}$,

35

where n = the number of samples. In our case, $n = 98$. Therefore, a threshold of 0.3 roughly corresponds to $3 - \sigma$ in the distribution ($3 \times 1/\sqrt{n-3}$).

2,460 such genes were found to satisfy this criterion. In order to evaluate the significance of the correlation coefficient of each gene with the ER level, a bootstrap technique was used to generate Monte-Carlo data that randomize the association between gene expression data of the samples and their categories. The distribution of correlation coefficients obtained from one Monte-Carlo trial is shown as a dashed line in FIG 5A. To estimate the significance of the 2,460 marker genes as a group, 10,000 Monte-Carlo runs were generated. The collection of 10,000 such Monte-Carlo trials forms the null hypothesis. The number of genes that satisfy the same criterion for Monte-Carlo data varies from run to run. The frequency distribution from 10,000 Monte-Carlo runs of the number of genes having correlation coefficients of >0.3 or <-0.3 is displayed in FIG. 5B. Both the mean and maximum value are much smaller than 2,460. Therefore, the significance of this gene group as the discriminating gene set between ER(+) and ER(-) samples is estimated to be greater than 99.99%.

2. Rank-ordering of candidate discriminating genes

In the second step, genes on the candidate list were rank-ordered based on the significance of each gene as a discriminating gene. The markers were rank-ordered either by amplitude of correlation, or by using a metric similar to a Fisher statistic:

$$t = \frac{(\langle x_1 \rangle - \langle x_2 \rangle)}{\sqrt{[\sigma_1^2(n_1 - 1) + \sigma_2^2(n_2 - 1)] / (n_1 + n_2 - 1) / (1/n_1 + 1/n_2)}} \quad \text{Equation (3)}$$

In Equation (3), $\langle x_1 \rangle$ is the error-weighted average of log ratio within the ER(-), and $\langle x_2 \rangle$ is the error-weighted average of log ratio within the ER(+) group. σ_1 is the variance of log ratio within the ER(-) group and n_1 is the number of samples that had valid measurements of log ratios. σ_2 is the variance of log ratio within the ER(+) group and n_2 is the number of samples that had valid measurements of log ratios. The t -value in Equation (3) represents the variance-compensated difference between two means. The confidence level of each gene in the candidate list was estimated with respect to a null hypothesis derived from the

actual data set using a bootstrap technique; that is, many artificial data sets were generated by randomizing the association between the clinical data and the gene expression data.

3. Optimization of the number of marker genes

The leave-one-out method was used for cross validation in order to optimize the discriminating genes. For a set of marker genes from the rank-ordered candidate list, a classifier was trained with 97 samples, and was used to predict the status of the remaining sample. The procedure was repeated for each of the samples in the pool, and the number of cases where the prediction for the one left out is wrong or correct was counted.

The above performance evaluation from leave-one-out cross validation was repeated by successively adding more marker genes from the candidate list. The performance as a function of the number of marker genes is shown in FIG. 6. The error rates for type 1 and type 2 errors varied with the number of marker genes used, but were both minimal while the number of the marker genes is around 550. Therefore, we consider this set of 550 genes is considered the optimal set of marker genes that can be used to classify breast cancer tumors into "ER-negative" group and "ER-positive" group. FIG. 7 shows the classification of patients as ER(+) or ER(-) based on this 550 marker set. FIG. 8 shows the correlation of each tumor to the ER-negative template verse the correlation of each tumor to the ER-positive template.

4. Classification based on marker genes

In the third step, a set of classifier parameters was calculated for each type of training data set based on either of the above ranking methods. A template for the ER(-)

group (\vec{Z}_1) was generated using the error-weighted log ratio average of the selected group of genes. Similarly, a template for ER(+) group (called \vec{Z}_2) was generated using the error-

weighted log ratio average of the selected group of genes. Two classifier parameters (P_1

and P_2) were defined based on either correlation or distance. P_1 measures the similarity

between one sample \vec{y} and the ER(-) template \vec{Z}_1 over this selected group of genes. P_2

measures the similarity between one sample \vec{y} and the ER(+) template \vec{Z}_2 over this selected

group of genes. The correlation P_i is defined as:

$$P_i = (\bar{z}_i \cdot \bar{y}) / (\|\bar{z}_i\| \cdot \|\bar{y}\|)$$

Equation (1)

5 A "leave-one-out" method was used to cross-validate the classifier built based on the marker genes. In this method, one sample was reserved for cross validation each time the classifier was trained. For the set of 550 optimal marker genes, the classifier was trained with 97 of the 98 samples, and the status of the remaining sample was predicted. This procedure was performed with each of the 98 patients. The number of cases where the prediction was wrong or correct was counted. It was further determined
10 that subsets of as few as ~50 of the 2,460 genes are able classify tumors as ER(+) or ER(-) nearly as well as using the total set.

In a small number of cases, there was disagreement between classification by the 550 marker set and a clinical classification. In comparing the microarray measured log ratio of expression for *ESR1* to the clinical binary decision (negative or positive) of ER
15 status for each patient, it was seen that the measured expression is consistent with the qualitative category of clinical measurements (mixture of two methods) for the majority of tumors. For example, two patients who were clinically diagnosed as ER(+) actually exhibited low expression of *ESR1* from microarray measurements and were classified as ER negative by 550 marker genes. Additionally, 3 patients who were clinically diagnosed as
20 ER(-) exhibited high expression of *ESR1* from microarray measurements and were classified as ER(+) by the same 550 marker genes. Statistically, however, microarray measured gene expression of *ESR1* correlates with the dominant patterns better than clinically determined ER status.

25 Example 3: Identification of Genetic Markers Distinguishing *BRCA1* Tumors From Sporadic Tumors in Estrogen Receptor (-) Patients

The *BRCA1* mutation is one of the major clinical categories in breast cancer tumors. It was determined that of tumors of 38 patients in the ER(-) group, 17 exhibited the *BRCA1* mutation, while 21 were sporadic tumors. A method was therefore developed that
30 enabled the differentiation of the 17 *BRCA1* mutation tumors from the 21 sporadic tumors in the ER(-) group.

1. Selection of candidate discriminating genes

In the first step, a set of candidate genes was identified based on the gene
35 expression patterns of these 38 samples. We first calculated the correlation between the

BRCA1-mutation category number and the expression ratio across all 38 samples for each individual gene by Equation (2). The distribution of the correlation coefficients is shown as a histogram defined by the solid line in FIG. 9A. We observed that, while the majority of genes do not correlate with *BRCA1* mutation status, a small group of genes correlated at significant levels. It is likely that genes with larger correlation coefficients would serve as reporters for discriminating tumors of *BRCA1* mutation carriers from sporadic tumors within the ER(-) group.

In order to evaluate the significance of each correlation coefficient with respect to a null hypothesis that such correlation coefficient could be found by chance, a bootstrap technique was used to generate Monte-Carlo data that randomizes the association between gene expression data of the samples and their categories. 10,000 such Monte-Carlo runs were generated as a control in order to estimate the significance of the marker genes as a group. A threshold of 0.35 in the absolute amplitude of correlation coefficients (either correlation or anti-correlation) was applied both to the real data and the Monte-Carlo data. Following this method, 430 genes were found to satisfy this criterion for the experimental data. The p-value of the significance, as measured against the 10,000 Monte-Carlo trials, is approximately 0.0048 (FIG. 9B). That is, the probability that this set of 430 genes contained useful information about *BRCA1*-like tumors vs sporadic tumors exceeds 99%.

2. Rank-ordering of candidate discriminating genes

In the second step, genes on the candidate list were rank-ordered based on the significance of each gene as a discriminating gene. Here, we used the absolute amplitude of correlation coefficients to rank order the marker genes.

3. Optimization of discriminating genes

In the third step, a subset of genes from the top of this rank-ordered list was used for classification. We defined a *BRCA1* group template (called \vec{Z}_1) by using the error-weighted log ratio average of the selected group of genes. Similarly, we defined a non-*BRCA1* group template (called \vec{Z}_2) by using the error-weighted log ratio average of the selected group of genes. Two classifier parameters (P1 and P2) were defined based on either correlation or distance. P1 measures the similarity between one sample \vec{Y} and the *BRCA1* template \vec{Z}_1 over this selected group of genes. P2 measures the similarity between

one sample \bar{y} and the non-*BRCA1* template \bar{z}_2 over this selected group of genes. For correlation, $P1$ and $P2$ were defined in the same way as in Equation (4).

The leave-one-out method was used for cross validation in order to optimize the discriminating genes as described in Example 2. For a set of marker genes from the rank-ordered candidate list, the classifier was trained with 37 samples the remaining one was predicted. The procedure was repeated for all the samples in the pool, and the number of cases where the prediction for the one left out is wrong or correct was counted.

To determine the number of markers constituting a viable subset, the above performance evaluation from leave-one-out cross validation was repeated by cumulatively adding more marker genes from the candidate list. The performance as a function of the number of marker genes is shown in FIG. 10. The error rates for type 1 (false negative) and type 2 (false positive) errors (Bendat & Piersol, RANDOM DATA ANALYSIS AND MEASUREMENT PROCEDURES, 2D ED., Wiley Interscience, p. 89) reached optimal ranges when the number of the marker genes is approximately 100. Therefore, a set of about 100 genes is considered to be the optimal set of marker genes that can be used to classify tumors in the ER(-) group as either *BRCA1*-related tumors or sporadic tumors.

The classification results using the optimal 100 genes are shown in FIGS. 11A and 11B. As shown in Figure 11A, the co-regulation patterns of the sporadic patients differ from those of the *BRCA1* patients primarily in the amplitude of regulation. Only one sporadic tumor was classified into the *BRCA1* group. Patients in the sporadic group are not necessarily *BRCA1* mutation negative; however, it is estimated that only approximately 5% of sporadic tumors are indeed *BRCA1*-mutation carriers.

Example 4: Identification of Genetic Markers Distinguishing Sporadic Tumor Patients with >5 Year Versus <5 Year Survival Times

78 tumors from sporadic breast cancer patients were used to explore prognostic predictors from gene expression data. Of the 78 samples in this sporadic breast cancer group, 44 samples were known clinically to have had no distant metastases within 5 years since the initial diagnosis ("no distant metastases group") and 34 samples had distant metastases within 5 years since the initial diagnosis ("distant metastases group"). A group of 231 markers, and optimally a group of 70 markers, was identified that allowed differentiation between these two groups.

1. Selection of candidate discriminating genes

In the first step, a set of candidate discriminating genes was identified based on gene expression data of these 78 samples. The correlation between the prognostic category number (distant metastases vs no distant metastases) and the logarithmic expression ratio across all samples for each individual gene was calculated using Equation (2). The distribution of the correlation coefficients is shown as a solid line in FIG. 12A. FIG. 12A also shows the result of one Monte-Carlo run as a dashed line. We observe that even though the majority of genes do not correlate with the prognostic categories, a small group of genes do correlate. It is likely that genes with larger correlation coefficients would be more useful as reporters for the prognosis of interest – distant metastases group and no distant metastases group.

In order to evaluate the significance of each correlation coefficient with respect to a null hypothesis that such correlation coefficient can be found by chance, we used a bootstrap technique to generate data from 10,000 Monte-Carlo runs as a control (FIG. 12B). We then selected genes that either have the correlation coefficient larger than 0.3 (“correlated genes”) or less than -0.3 (“anti-correlated genes”). The same selection criterion was applied both to the real data and the Monte-Carlo data. Using this comparison, 231 markers from the experimental data were identified that satisfy this criterion. The probability of this gene set for discriminating patients between the distant metastases group and the no distant metastases group being chosen by random fluctuation is approximately 0.003.

2. Rank-ordering of candidate discriminating genes

In the second step, genes on the candidate list were rank-ordered based on the significance of each gene as a discriminating gene. Specifically, a metric similar to a “Fisher” statistic, defined in Equation (3), was used for the purpose of rank ordering. The confidence level of each gene in the candidate list was estimated with respect to a null hypothesis derived from the actual data set using the bootstrap technique. Genes in the candidate list can also be ranked by the amplitude of correlation coefficients.

3. Optimization of discriminating genes

In the third step, a subset of 5 genes from the top of this rank-ordered list was selected to use as discriminating genes to classify 78 tumors into a “distant metastases group” or a “no distant metastases group”. The leave-one-out method was used for cross validation. Specifically, 77 samples defined a classifier based on the set of selected

discriminating genes, and these were used to predict the remaining sample. This procedure was repeated so that each of the 78 samples was predicted. The number of cases in which predictions were correct or incorrect were counted. The performance of the classifier was measured by the error rates of type 1 and type 2 for this selected gene set.

- 5 We repeated the above performance evaluation procedure, adding 5 more marker genes each time from the top of the candidate list, until all 231 genes were used. As shown in FIG. 13, the number of mis-predictions of type 1 and type 2 errors change dramatically with the number of marker genes employed. The combined error rate reached a minimum when 70 marker genes from the top of our candidate list were used. Therefore, 10 this set of 70 genes is the optimal, preferred set of marker genes useful for the classification of sporadic tumor patients into either the distant metastases or no distant metastases group. Fewer or more markers also act as predictors, but are less efficient, either because of higher error rates, or the introduction of statistical noise.

15 4. Reoccurrence probability curves

The prognostic classification of 78 patients with sporadic breast cancer tumors into two distinct subgroups was predicted based on their expression of the 70 optimal marker genes (FIGS. 14 and 15).

- 20 To evaluate the prognostic classification of sporadic patients, we predicted the outcome of each patient by a classifier trained by the remaining 77 patients based on the 70 optimal marker genes. FIG. 16 plots the distant metastases probability as a function of the time since initial diagnosis for the two predicted groups. The difference between these two reoccurrence curves is significant. Using the χ^2 test (S-PLUS 2000 Guide to Statistics, vol. 2, MathSoft, p. 44), the p-value is estimated to be $\sim 10^{-9}$. The distant metastases 25 probability as a function of the time since initial diagnosis was also compared between ER(+) and ER(-) individuals (FIG. 17), PR(+) and PR(-) individuals (FIG. 18), and between individuals with different tumor grades (FIGS. 19A, 19B). In comparison, the p-values for the differences between two prognostic groups based on clinical data are much less significant than that based on gene expression data, ranging from 10^{-3} to 1.

- 30 To parameterize the reoccurrence probability as a function of time since initial diagnosis, the curve was fitted to one type of survival model – “normal”:

$$P = \alpha \times \exp\left(-t^2/\tau^2\right) \quad (4)$$

- For fixed $\alpha = 1$, we found that $\tau = 125$ months for patients in the no distant metastases group 35 and $\tau = 36$ months for patients in the distant metastases group. Using tumor grades, we

found $\tau = 100$ months for patients with tumor grades 1 and 2 and $\tau = 60$ for patients with tumor grade 3. It is accepted clinical practice that tumor grades are the best available prognostic predictor. However, the difference between the two prognostic groups classified based on 70 marker genes is much more significant than those classified by the best
5 available clinical information.

5. Prognostic prediction for 19 independent sporadic tumors

To confirm the proposed prognostic classification method and to ensure the reproducibility, robustness, and predicting power of the 70 optimal prognostic marker
10 genes, we applied the same classifier to 19 independent tumor samples from sporadic breast cancer patients, prepared separately at The Netherlands Cancer Institute (NKI). The same reference pool was used.

The classification results of 19 independent sporadic tumors are shown in Figure 20. FIG. 20A shows the log ratio of expression regulation of the same 70 optimum
15 marker genes. Based on our classifier model, we expected the misclassification of $19 \cdot (6+7)/78 = 3.2$ tumors. Consistently, $(1+3) = 4$ of 19 tumors were misclassified.

6. Clinical parameters as a group vs. microarray data – Results of logistic regression

20 In the previous section, the predictive power of each individual clinical parameter was compared with that of the expression data. However, it is more meaningful to combine all the clinical parameters as a group, and then compare them to the expression data. This requires multi-variant modeling; the method chosen was logistic regression. Such an approach also demonstrates how much improvement the microarray approach adds
25 to the results of the clinical data.

The clinical parameters used for the multi-variant modeling were: (1) tumor grade; (2) ER status; (3) presence or absence of the progesterone receptor (PR); (4) tumor size; (5) patient age; and (6) presence or absence of angiogenesis. For the microarray data, two correlation coefficients were used. One is the correlation to the mean of the good
30 prognosis group (C1) and the other is the correlation to the mean of the bad prognosis group (C2). When calculating the correlation coefficients for a given patient, this patient is excluded from either of the two means.

The logistic regression optimizes the coefficient of each input parameter to best predict the outcome of each patient. One way to judge the predictive power of each
35 input parameter is by how much deviance (similar to Chi-square in the linear regression, see

for example, Hosmer & Lemeshow, APPLIED LOGISTIC REGRESSION, John Wiley & Sons, (2000)) the parameter accounts for. The best predictor should account for most of the deviance. To fairly assess the predictive power, each parameter was modeled independently. The microarray parameters explain most of the deviance, and hence are
5 powerful predictors.

The clinical parameters, and the two microarray parameters, were then monitored as a group. The total deviance explained by the six clinical parameters was 31.5, and total deviance explained by the microarray parameters was 39.4. However, when the clinical data was modeled first, and the two microarray parameters added, the final deviance
10 accounted for is 57.0.

The logistic regression computes the likelihood that a patient belongs to the good or poor prognostic group. FIGS. 21A and 21B show the sensitivity vs. (1-specificity). The plots were generated by varying the threshold on the model predicted likelihood. The curve which goes through the top left corner is the best (high sensitivity with high
15 specificity). The microarray outperformed the clinical data by a large margin. For example, at a fixed sensitivity of around 80%, the specificity was ~80% from the microarray data, and ~65% from the clinical data for the good prognosis group. For the poor prognosis group, the corresponding specificities were ~80% and ~70%, again at a fixed sensitivity of 80%. Combining the microarray data with the clinical data further improved the results.
20 The result can also be displayed as the total error rate as the function of the threshold in FIG. 21C. At all possible thresholds, the error rate from the microarray was always smaller than that from the clinical data. By adding the microarray data to the clinical data, the error rate is further reduced, as one can see in Figure 21C.

Odds ratio tables can be created from the prediction of the logistic
25 regression. The probability of a patient being in the good prognosis group is calculated by the logistic regression based on different combinations of input parameters (clinical and/or microarray). Patients are divided into the following four groups according to the prediction and the true outcome: (1) predicted good and truly good, (2) predicted good but truly poor, (3) predicted poor but truly good, (4) predicted poor and truly poor. Groups (1) & (4)
30 represent correct predictions, while groups (2) & (3) represent mis-predictions. The division for the prediction is set at probability of 50%, although other thresholds can be used. The results are listed in Table 7. It is clear from Table 7 that microarray profiling (Table 7.3 & 7.10) outperforms any single clinical data (Table 7.4-7.9) and the combination of the clinical data (Table 7.2). Adding the micro-array profiling in addition to the clinical
35 data give the best results (Table 7.1).

For microarray profiling, one can also make a similar table (Table 7.11) without using logistic regression. In this case, the prediction was simply based on C1-C2 (greater than 0 means good prognosis, less than 0 mean bad prognosis).

| | | | |
|----|---|----------------|----------------|
| 5 | Table 7.1 Prediction by clinical+microarray | | |
| | | Predicted good | Predicted poor |
| | true good | 39 | 5 |
| | true poor | 4 | 30 |
| | Table 7.2 Prediction by clinical alone | | |
| | | Predicted good | Predicted poor |
| 10 | true good | 34 | 10 |
| | true poor | 12 | 22 |
| | Table 7.3 Prediction by microarray | | |
| | | predicted good | Predicted poor |
| | true good | 39 | 5 |
| | true poor | 10 | 24 |
| 15 | Table 7.4 Prediction by grade | | |
| | | Predicted good | Predicted poor |
| | true good | 23 | 21 |
| | true poor | 5 | 29 |
| | Table 7.5 Prediction by ER | | |
| | | Predicted good | Predicted poor |
| 20 | true good | 35 | 9 |
| | true poor | 21 | 13 |
| | Table 7.6 Prediction by PR | | |
| | | Predicted good | Predicted poor |
| | true good | 35 | 9 |
| | true poor | 18 | 16 |
| | Table 7.7 Prediction by size | | |
| 25 | | Predicted good | Predicted poor |
| | true good | 35 | 9 |
| | true poor | 13 | 21 |
| | Table 7.8 Prediction by age | | |
| | | Predicted good | Predicted poor |
| 30 | true good | 33 | 11 |
| | true poor | 15 | 19 |
| | Table 7.9 Prediction by angioinvasion | | |
| | | Predicted good | Predicted poor |
| | true good | 37 | 7 |
| | true poor | 19 | 15 |
| | Table 7.10 Prediction by dC (C1-C2) | | |
| 35 | | Predicted good | Predicted poor |
| | true good | 36 | 8 |
| | true poor | 6 | 28 |

| Table 7.11 No logistic regression, simply judged by C1-C2 | | |
|---|----------------|----------------|
| | Predicted good | Predicted poor |
| true good | 37 | 7 |
| true poor | 6 | 28 |

5

Example 5. Concept of mini-array for diagnosis purposes.

All genes on the marker gene list for the purpose of diagnosis and prognosis can be synthesized on a small-scale microarray using ink-jet technology. A microarray with genes for diagnosis and prognosis can respectively or collectively be made. Each gene on the list is represented by single or multiple oligonucleotide probes, depending on its sequence uniqueness across the genome. This custom designed mini-array, in combination with sample preparation protocol, can be used as a diagnostic/prognostic kit in clinics.

Example 6. Biological Significance of diagnostic marker genes

The public domain was searched for the available functional annotations for the 430 marker genes for *BRCA1* diagnosis in Table 3. The 430 diagnostic genes in Table 3 can be divided into two groups: (1) 196 genes whose expressions are highly expressed in *BRCA1*-like group; and (2) 234 genes whose expression are highly expressed sporadic group. Of the 196 *BRCA1* group genes, 94 are annotated. Of the 234 sporadic group genes, 100 are annotated. The terms "T-cell", "B-cell" or "immunoglobulin" are involved in 13 of the 94 annotated genes, and in 1 of the 100 annotated genes, respectively. Of 24,479 genes represented on the microarrays, there are 7,586 genes with annotations to date. "T-cell", "B-cell" and "immunoglobulin" are found in 207 of these 7,586 genes. Given this, the p-value of the 13 "T-cell", "B-cell" or "immunoglobulin" genes in the *BRCA1* group is very significant (p-value = 1.1×10^{-6}). In comparison, the observation of 1 gene relating to "T-cell", "B-cell", or "immunoglobulin" in the sporadic group is not significant (p-value = 0.18).

The observation that *BRCA1* patients have highly expressed lymphocyte (T-cell and B-cell) genes agrees with what has been seen from pathology that *BRCA1* breast tumor has more frequently associated with high lymphocytic infiltration than sporadic cases (Chappuis *et al.*, 2000, *Semin Surg Oncol* 18:287-295).

Example 7. Biological significance of prognosis marker genes

A search was performed for available functional annotations for the 231 prognosis marker genes (Table 5). The markers fall into two groups: (1) 156 markers

whose expressions are highly expressed in poor prognostic group; and (2) 75 genes whose expression are highly expressed in good prognostic group. Of the 156 markers, 72 genes are annotated; of the 75 genes, 28 genes are annotated.

Twelve of the 72 markers, but none of the 28 markers, are, or are associated
5 with, kinases. In contrast, of the 7,586 genes on the microarray having annotations to date, only 471 involve kinases. On this basis, the p-value that twelve kinase-related markers in the poor prognostic group is significant (p-value = 0.001). Kinases are important regulators of intracellular signal transduction pathways mediating cell proliferation, differentiation and apoptosis. Their activity is normally tightly controlled and regulated. Overexpression of
10 certain kinases is well known involving in oncogenesis, such as vascular endothelial growth factor receptor1 (VEGFR1 or FLT1), a tyrosine kinase in the poor prognosis group, which plays a very important role in tumor angiogenesis. Interestingly, vascular endothelial growth factor (VEGF), VEGFR's ligand, is also found in the prognosis group, which means both ligand and receptor are upregulated in poor prognostic individuals by an unknown
15 mechanism.

Likewise, 16 of the 72 markers, and only two of the 28 markers, are, or are associated with, ATP-binding or GTP-binding proteins. In contrast, of the 7,586 genes on the microarray having annotations to date, only 714 and 153 involve ATP-binding and GTP-binding, respectively. On this basis, the p-value that 16 GTP- or ATP-binding-related
20 markers in the poor prognosis group is significant (p-value 0.001 and 0.0038). Thus, the kinase- and ATP- or GTP-binding-related markers within the 72 markers can be used as prognostic indicators.

Cancer is characterized by deregulated cell proliferation. On the simplest level, this requires division of the cell or mitosis. By keyword searching, we found "cell
25 division" or "mitosis" included in the annotations of 7 genes respectively in the 72 annotated markers from the 156 poor prognosis markers, but in none for the 28 annotated genes from 75 good prognosis markers. Of the 7,586 microarray markers with annotations, "cell division" is found in 62 annotations and "mitosis" is found in 37 annotations. Based on these findings, the p-value that seven cell division- or mitosis-related markers are found
30 in the poor prognosis group is estimated to be highly significant (p-value = 3.5×10^{-5}). In comparison, the absence of cell division- or mitosis-related markers in the good prognosis group is not significant (p-value = 0.69). Thus, the seven cell division- or mitosis-related markers may be used as markers for poor prognosis.

Example 8: Construction of an artificial reference pool.

The reference pool for expression profiling in the above Examples was made by using equal amount of cRNAs from each individual patient in the sporadic group. In order to have a reliable, easy-to-made, and large amount of reference pool, a reference pool
5 for breast cancer diagnosis and prognosis can be constructed using synthetic nucleic acid representing, or derived from, each marker gene. Expression of marker genes for individual patient sample is monitored only against the reference pool, not a pool derived from other patients.

To make the reference pool, 60-mer oligonucleotides are synthesized
10 according to 60-mer ink-jet array probe sequence for each diagnostic/prognostic reporter genes, then double-stranded and cloned into pBluescript SK- vector (Stratagene, La Jolla, CA), adjacent to the T7 promoter sequence. Individual clones are isolated, and the sequences of their inserts are verified by DNA sequencing. To generate synthetic RNAs, clones are linearized with *EcoRI* and a T7 *in vitro* transcription (IVT) reaction is performed
15 according to the MegaScript kit (Ambion, Austin, TX). IVT is followed by DNase treatment of the product. Synthetic RNAs are purified on RNeasy columns (Qiagen, Valencia, CA). These synthetic RNAs are transcribed, amplified, labeled, and mixed together to make the reference pool. The abundance of those synthetic RNAs are adjusted to approximate the abundance of the corresponding marker-derived transcripts in the real
20 tumor pool.

Example 9: Use of single-channel data and a sample pool represented by stored values.

1. Creation of a reference pool of stored values ("mathematical sample pool")

The use of ratio-based data used in Examples 1-7, above, requires a physical
25 reference sample. In the above Examples, a pool of sporadic tumor sample was used as the reference. Use of such a reference, while enabling robust prognostic and diagnostic predictions, can be problematic because the pool is typically a limited resource. A classifier method was therefore developed that does not require a physical sample pool, making application of this predictive and diagnostic technique much simpler in clinical applications.

30 To test whether single-channel data could be used, the following procedure was developed. First, the single channel intensity data for the 70 optimal genes, described in Example 4, from the 78 sporadic training samples, described in the Materials and Methods, was selected from the sporadic sample vs. tumor pool hybridization data. The 78 samples consisted of 44 samples from patients having a good prognosis and 34 samples
35 from patients having a poor prognosis. Next, the hybridization intensities for these samples

were normalized by dividing by the median intensity of all the biological spots on the same microarray. Where multiple microarrays per sample were used, the average was taken across all of the microarrays. A log transform was performed on the intensity data for each of the 70 genes, or for the average intensity for each of the 70 genes where more than one
5 microarray is hybridized, and a mean log intensity for each gene across the 78 sporadic samples was calculated. For each sample, the mean log intensities thus calculated were subtracted from the individual sample log intensity. This figure, the mean subtracted $\log(\text{intensity})$ was then treated as the two color $\log(\text{ratio})$ for the classifier by substitution into Equation (5). For new samples, the mean log intensity is subtracted in the same
10 manner as noted above, and a mean subtracted $\log(\text{intensity})$ calculated.

The creation of a set of mean log intensities for each gene hybridized creates a “mathematical sample pool” that replaces the quantity-limited “material sample pool.” This mathematical sample pool can then be applied to any sample, including samples in hand and ones to be collected in the future. This “mathematical sample pool” can be
15 updated as more samples become available.

2. Results

To demonstrate that the mathematical sample pool performs a function equivalent to the sample reference pool, the mean-subtracted- $\log(\text{intensity})$ (single channel
20 data, relative to the mathematical pool) vs. the $\log(\text{ratio})$ (hybridizations, relative to the sample pool) was plotted for the 70 optimal reporter genes across the 78 sporadic samples, as shown in FIG. 22. The ratio and single-channel quantities are highly correlated, indicating both have the capability to report relative changes in gene expression. A classifier was then constructed using the mean-subtracted- $\log(\text{intensity})$ following exactly
25 the same procedure as was followed using the ratio data, as in Example 4.

As shown in FIGS. 23A and 23B, single-channel data was successful at classifying samples based on gene expression patterns. FIG. 23A shows samples grouped according to prognosis using single-channel hybridization data. The white line separates samples from patients classified as having poor prognoses (below) and good prognoses
30 (above). FIG. 23B plots each sample as its expression data correlates with the good (open circles) or poor (filled squares) prognosis classifier parameter. Using the “leave-one-out” cross validation method, the classifier predicted 10 false positives out of 44 samples from patients having a good prognosis, and 6 false negatives out of 34 samples from patients having a poor prognosis, where a poor prognosis is considered a “positive.” This outcome
35

is comparable to the use of the ratio-based classifier, which predicted 7 out of 44, and 6 out of 34, respectively.

In clinical applications, it is greatly preferable to have few false positives, which results in fewer under-treated patients. To conform the results to this preference, a classifier was constructed by ranking the patient sample according to its coefficient of correlation to the "good prognosis" template, and chose a threshold for this correlation coefficient to allow approximately 10% false negatives, *i.e.*, classification of a sample from a patient with poor prognosis as one from a patient with a good prognosis. Out of the 34 poor prognosis samples used herein, this represents a tolerance of 3 out of 34 poor prognosis patients classified incorrectly. This tolerance limit corresponds to a threshold 0.2727 coefficient of correlation to the "good prognosis" template. Results using this threshold are shown in FIGS. 24A and 24B. FIG. 24A shows single-channel hybridization data for samples ranked according to the coefficients of correlation with the good prognosis classifier; samples classified as "good prognosis" lie above the white line, and those classified as "poor prognosis" lie below. FIG. 24B shows a scatterplot of sample correlation coefficients, with three incorrectly classified samples lying to the right of the threshold correlation coefficient value. Using this threshold, the classifier had a false positive rate of 15 out of the 44 good prognosis samples. This result is not very different compared to the error rate of 12 out of 44 for the ratio based classifier.

In summary, the 70 reporter genes carry robust information about prognosis; the single channel data can predict the tumor outcome almost as well as the ratio based data, while being more convenient in a clinical setting.

7. REFERENCES CITED

All references cited herein are incorporated herein by reference in their entirety and for all purposes to the same extent as if each individual publication or patent or patent application was specifically and individually indicated to be incorporated by
5 reference in its entirety for all purposes.

Many modifications and variations of the present invention can be made without departing from its spirit and scope, as will be apparent to those skilled in the art. The specific embodiments described herein are offered by way of example only, and the invention is to be limited only by the terms of the appended claims along with the full scope
10 of equivalents to which such claims are entitled.

15

20

25

30

35

What is claimed is:

1. A method for classifying a cell sample as ER(+) or ER(-) comprising detecting a difference in the expression by said cell sample of a first plurality of genes relative to a control, said first plurality of genes consisting of at least 5 of the genes
5 corresponding to the markers listed in Table 1.
2. The method of claim 1, wherein said plurality consists of at least 50 of the genes corresponding to the markers listed in Table 1.
- 10 3. The method of claim 1, wherein said plurality consists of at least 100 of the genes corresponding to the markers listed in Table 1.
4. The method of claim 1, wherein said plurality consists of at least 200 of the genes corresponding to the markers listed in Table 1.
- 15 5. The method of claim 1, wherein said plurality consists of at least 500 of the genes corresponding to the markers listed in Table 1.
6. The method of claim 1, wherein said plurality consists of at least
20 1000 of the genes corresponding to the markers listed in Table 1.
7. The method of claim 1, wherein said plurality consists of each of the genes corresponding to the 2,460 markers listed in Table 2.
- 25 8. The method of claim 1, wherein said plurality consists of the 550 gene markers listed in Table 2.
9. The method of claim 1, wherein said control comprises nucleic acids derived from a pool of tumors from individual sporadic patients.
- 30 10. The method of claim 1, wherein said detecting comprises the steps of
(a) generating an ER(+) template by hybridization of nucleic acids derived from a plurality of ER(+) patients within a plurality of sporadic patients against nucleic acids derived from a pool of tumors from individual sporadic patients;
- 35

- (b) generating an ER(-) template by hybridization of nucleic acids derived from a plurality of ER(-) patients within said plurality of sporadic patients against nucleic acids derived from said pool of tumors from individual sporadic patients within said plurality;
- 5 (c) hybridizing an nucleic acids derived from an individual sample against said pool; and
- (d) determining the similarity of marker gene expression in the individual sample to the ER(+) template and the ER(-) template, wherein if said expression is more similar to the ER(+) template, the sample is classified as ER(+), and if said
- 10 expression is more similar to the ER(-) template, the sample is classified as ER(-).
11. A method for classifying a cell sample as *BRCA1*-related or sporadic, comprising detecting a difference in the expression of a first plurality of genes relative to a control, said first plurality of genes consisting of at least 5 of the genes
- 15 corresponding to the markers listed in Table 3.
12. The method of claim 11, wherein said plurality consists of at least 50 of the genes corresponding to the markers listed in Table 3.
- 20 13. The method of claim 11, wherein said plurality consists of at least 100 of the genes corresponding to the markers listed in Table 3.
14. The method of claim 11, wherein said plurality consists of at least 200 of the genes corresponding to the markers listed in Table 3.
- 25 15. The method of claim 11, wherein said plurality consists of each of the genes corresponding to the 430 markers listed in Table 3.
16. The method of claim 11, wherein said plurality consists of each of the
- 30 genes corresponding to the 100 markers listed in Table 4.
17. The method of claim 11, wherein said control comprises nucleic acids derived from a pool of tumors from individual sporadic patients.

18. The method of claim 11, wherein said detecting comprises the steps of

- (a) generating a *BRCA1* template by hybridization of nucleic acids derived from a plurality of *BRCA1* patients within a plurality of ER(-) patients against
5 nucleic acids derived from a pool of tumors;
- (b) generating a sporadic template by hybridization of nucleic acids derived from a plurality of sporadic patients within said plurality of ER(-) patients against nucleic acids derived from said pool of tumors;
- (c) hybridizing nucleic acids derived from an individual sample against
10 said pool; and
- (d) determining the similarity of marker gene expression in the individual sample to the *BRCA1* template and the sporadic template, wherein if said expression is more similar to the *BRCA1* template, the sample is classified as *BRCA1*, and if said expression is more similar to the sporadic template, the sample is classified as sporadic.

15

19. A method for classifying an individual as having a good prognosis (no distant metastases within five years of initial diagnosis) or a poor prognosis (distant metastases within five years of initial diagnosis), comprising detecting a difference in the expression of a first plurality of genes in a cell sample taken from the individual relative to a
20 control, said first plurality of genes consisting of at least 5 of the genes corresponding to the markers listed in Table 5.

20. The method of claim 19, wherein said plurality consists of at least 20 of the genes corresponding to the markers listed in Table 5.

25

21. The method of claim 19, wherein said plurality consists of at least 100 of the genes corresponding to the markers listed in Table 5.

22. The method of claim 19, wherein said plurality consists of at least
30 150 of the genes corresponding to the markers listed in Table 5.

23. The method of claim 19, wherein said plurality consists of each of the genes corresponding to the 231 markers listed in Table 5.

35

24. The method of claim 19, wherein said plurality consists of the 70 gene markers listed in Table 6.

25. The method of claim 1, wherein said control comprises nucleic acids
5 derived from a pool of tumors from individual sporadic patients.

26. The method of claim 19, wherein said detecting comprises the steps of:

(a) generating a good prognosis template by hybridization of nucleic
10 acids derived from a plurality of good prognosis patients against nucleic acids derived from a pool of tumors from individual patients;

(b) generating a poor prognosis template by hybridization of nucleic acids derived from a plurality of poor prognosis patients against nucleic acids derived from said pool of tumors from said plurality of individual patients;

15 (c) hybridizing an nucleic acids derived from and individual sample against said pool; and

(d) determining the similarity of marker gene expression in the individual sample to the good prognosis template and the poor prognosis template, wherein if said expression is more similar to the good prognosis template, the sample is classified as
20 having a good prognosis, and if said expression is more similar to the poor prognosis template, the sample is classified as having a poor prognosis.

27. The method of claim 1, wherein the cell sample is additionally classified as *BRCAl*-related or sporadic by detecting a difference in the expression of a
25 second plurality of genes in a cell sample taken from the individual relative to a control, said second plurality of genes consisting of at least 5 of the genes corresponding to the markers listed in Table 3 or Table 4.

28. The method of claim 1, wherein the cell sample is additionally
30 classified as taken from a patient with a good prognosis or a poor prognosis by detecting a difference in the expression of a second plurality of genes in a cell sample taken from the individual relative to a control, said second plurality of genes consisting of at least 5 of the genes corresponding to the markers listed in Table 5.

35

29. The method of claim 11, wherein the cell sample is additionally classified as taken from a patient with a good prognosis or a poor prognosis by detecting a difference in the expression of a second plurality of genes in a cell sample taken from the individual relative to a control, said second plurality of genes consisting of at least 20 of the
5 genes corresponding to the markers listed in Table 5.

30. The method of claim 11, wherein the cell sample is additionally classified as ER(+) or ER(-) by detecting a difference in the expression of a second plurality of genes in a cell sample taken from the individual relative to a control, said second
10 plurality of genes consisting of at least 5 of the genes corresponding to the markers listed in Table 1.

31. The method of claim 19, wherein the cell sample is additionally classified as ER(+) or ER(-) by detecting a difference in the expression of a second plurality
15 of genes in a cell sample taken from the individual relative to a control, said second plurality of genes consisting of at least 5 of the genes corresponding to the markers listed in Table 1.

32. The method of claim 19, wherein the cell sample is additionally
20 classified as *BRCA1* or sporadic by detecting a difference in the expression of a second plurality of genes in a cell sample taken from the individual relative to a control, said second plurality of genes consisting of at least 5 of the genes corresponding to the markers listed in Table 3.

25 33. A method for classifying a sample as ER(+) or ER(-) by calculating the similarity between the expression of at least 5 of the markers listed in Table 1 in the sample to the expression of the same markers in an ER(-) nucleic acid pool and an ER(+) nucleic acid pool, comprising the steps of:

(a) labeling nucleic acids derived from a sample, with a first fluorophore
30 to obtain a first pool of fluorophore-labeled nucleic acids;

(b) labeling with a second fluorophore a first pool of nucleic acids derived from two or more ER(+) samples, and a second pool of nucleic acids derived from two or more ER(-) samples:

(c) contacting said first fluorophore-labeled nucleic acid and said first
35 pool of second fluorophore-labeled nucleic acid with a first microarray under conditions

such that hybridization can occur, and contacting said first fluorophore-labeled nucleic acid and said second pool of second fluorophore-labeled nucleic acid with a second microarray under conditions such that hybridization can occur, wherein said first microarray and said second microarray are similar to each other, exact replicas of each other, or are identical,
5 detecting at each of a plurality of discrete loci on the first microarray a first fluorescent emission signal from said first fluorophore-labeled nucleic acid and a second fluorescent emission signal from said first pool of second fluorophore-labeled genetic matter that is bound to said first microarray under said conditions, and detecting at each of the marker loci on said second microarray said first fluorescent emission signal from said first fluorophore-
10 labeled nucleic acid and a third fluorescent emission signal from said second pool of second fluorophore-labeled nucleic acid;

(d) determining the similarity of the sample to the ER(-) and ER(+) pools by comparing said first fluorescence emission signals and said second fluorescence emission signals, and said first emission signals and said third fluorescence emission
15 signals; and

(e) classifying the sample as ER(+) where the first fluorescence emission signals are more similar to said second fluorescence emission signals than to said third fluorescent emission signals, and classifying the sample as ER(-) where the first fluorescence emission signals are more similar to said third fluorescence emission signals
20 than to said second fluorescent emission signals.

34. The method of claim 33, wherein said similarity is calculated by determining a first sum of the differences of expression levels for each marker between said first fluorophore-labeled nucleic acid and said first pool of second fluorophore-labeled
25 nucleic acid, and a second sum of the differences of expression levels for each marker between said first fluorophore-labeled nucleic acid and said second pool of second fluorophore-labeled nucleic acid, wherein if said first sum is greater than said second sum, the sample is classified as ER(-), and if said second sum is greater than said first sum, the sample is classified as ER(+).

30

35. The method of claim 33, wherein said similarity is calculated by computing a first classifier parameter P_1 between an ER(+) template and the expression of said markers in said sample, and a second classifier parameter P_2 between an ER(-) template and the expression of said markers in said sample, wherein said P_1 and P_2 are calculated
35 according to the formula:

$$P_i = (\vec{z}_i \bullet \vec{y}) / (\|\vec{z}_i\| \cdot \|\vec{y}\|),$$

wherein \vec{z}_1 and \vec{z}_2 are ER(+) and ER(-) templates, respectively, and are calculated by
 5 averaging said second fluorescence emission signal for each of said markers in said first pool of second fluorophore-labeled nucleic acid and said third fluorescence emission signal for each of said markers in said second pool of second fluorophore-labeled nucleic acid, respectively, and wherein \vec{y} is said first fluorescence emission signal of each of said
 10 markers in the sample to be classified as ER(+) or ER(-), wherein the expression of the markers in the sample is similar to ER(-) if $P_1 < P_2$, and similar to ER(+) if $P_1 > P_2$.

36. A method for determining a set of marker genes whose expression is associated with a particular phenotype, comprising the steps of:

- 15 (a) selecting phenotype having two or more phenotype categories;
 (b) identifying a plurality of genes wherein the expression of said genes is correlated or anticorrelated with one of the phenotype categories, and wherein the correlation coefficient for each gene is calculated according to the equation

20 $\rho = (\vec{c} \bullet \vec{r}) / (\|\vec{c}\| \cdot \|\vec{r}\|)$, wherein \vec{c} is a number representing said phenotype category and \vec{r} is the logarithmic expression ratio across all the samples for each individual gene, wherein if the correlation coefficient has an absolute value of 0.3 or greater, said expression of said gene is associated with the phenotype category,

wherein said plurality of genes is a set of marker genes whose expression is
 25 associated with a particular phenotype.

37. The method of claim 36, wherein said set of marker genes is validated by:

- (a) using a statistical method to randomize the association between said
 30 marker genes and said phenotype category, thereby creating a control correlation coefficient for each marker gene;
 (b) repeating step (a) one hundred or more times to develop a frequency distribution of said control correlation coefficients for each marker gene;
 (c) determining the number of marker genes having a control correlation
 35 coefficient of 0.3 or above, thereby creating a control marker gene set; and

(d) comparing the number of control marker genes so identified to the number of marker genes, wherein if the p value of the difference between the number of marker genes and the number of control genes is less than a threshold, said set of marker genes is validated.

5

38. The method of claim 36, wherein said set of marker genes is optimized by the method comprising:

(a) rank-ordering the genes by amplitude of correlation or by significance of the correlation coefficients to create a rank-ordered list, and

10

(b) selecting an arbitrary number n of marker genes from the top of the rank-ordered list.

39. The method of claim 38, wherein said set of marker genes is further optimized by the method comprising:

15

(a) calculating an error rate for said arbitrary number n of marker genes;

(b) increasing by 1 the number of genes selected from the top of the rank-ordered list;

(c) calculating an error rate for said number of genes selected from the top of the rank-ordered list;

20

(d) repeating steps (b) and (c) until said number of genes selected from the top of the rank-ordered list includes all genes included in said rank ordered list, and

(e) identifying said number of genes selected from the top of the rank-ordered list for which the error rate is smallest,

wherein said set of marker genes is optimized when the error rate is the smallest.

25

40. A method for assigning a person to one of a plurality of categories in a clinical trial, comprising determining for each said person the level of expression of at least five of the prognosis markers listed in Table 6, determining therefrom whether the person has an expression pattern that correlates with a good prognosis or a poor prognosis, and assigning said person to one category in a clinical trial if said person is determined to have a good prognosis, and a different category if that person is determined to have a poor prognosis.

35

41. A method of classifying a first cell or organism as having one of at least two different phenotypes, said at least two different phenotypes comprising a first phenotype and a second phenotype, said method comprising:

5 (a) comparing the level of expression of each of a plurality of genes in a first sample from the first cell or organism to the level of expression of each of said genes, respectively, in a pooled sample from a plurality of cells or organisms, said plurality of cells or organisms comprising different cells or organisms exhibiting said at least two different phenotypes, respectively, to produce a first compared value;

10 (b) comparing said first compared value to a second compared value, wherein said second compared value is the product of a method comprising comparing the level of expression of each of said genes in a sample from a cell or organism characterized as having said first phenotype to the level of expression of each of said genes, respectively, in said pooled sample;

15 (c) comparing said first compared value to a third compared value, wherein said third compared value is the product of a method comprising comparing the level of expression of each of said genes in a sample from a cell or organism characterized as having said second phenotype to the level of expression of each of said genes, respectively, in said pooled sample,

20 (d) optionally carrying out one or more times a step of comparing said first compared value to one or more additional compared values, respectively, each additional compared value being the product of a method comprising comparing the level of expression of each of said genes in a sample from a cell or organism characterized as having a phenotype different from said first and second phenotypes but included among said at least two different phenotypes, to the level of expression of each of said genes, respectively,
25 in said pooled sample; and

(e) determining to which of said second, third and, if present, one or more additional compared values, said first compared value is most similar;

30 wherein said first cell or organism is determined to have the phenotype of the cell or organism used to produce said compared value most similar to said first compared value.

42. The method of claim 40, wherein said compared values are each ratios of the levels of expression of each of said genes.

43. The method of claim 40, wherein each of said levels of expression of each of said genes in said pooled sample are normalized prior to any of said comparing steps.

5 44. The method of claim 42 wherein normalizing said levels of expression is carried out by dividing each of said levels of expression by the median or mean level of expression of each of said genes or dividing by the mean or median level of expression of one or more housekeeping genes in said pooled sample.

10 45. The method of claim 42 wherein said normalized levels of expression are subjected to a log transform and said comparing steps comprise subtracting said log transform from the log of said levels of expression of each of said genes in said sample from said cell or organism.

15 46. The method of claim 40, wherein said at least two different phenotypes are different stages of a disease or disorder.

47. The method of claim 40, wherein said at least two different phenotypes are different prognoses of a disease or disorder.

20 48. The method of claim 40, wherein said levels of expression of each of said genes, respectively, in said pooled sample or said levels of expression of each of said genes in a sample from said cell or organism characterized as having said first phenotype, said second phenotype, or said phenotype different from said first and second phenotypes, respectively, are stored on a computer.

49. A microarray comprising at least 5 markers derived from any one of Tables 1-6, wherein at least 50% of the probes on the microarray are present in any one of Tables 1-6.

30 50. The microarray of claim 48, wherein at least 70% of the probes on the microarray are present in any one of Tables 1-6.

51. The microarray of claim 48, wherein at least 80% of the probes on
35 the microarray are present in any one of Tables 1-6.

52. The microarray of claim 48, wherein at least 90% of the probes on the microarray are present in any one of Tables 1-6.

53. The microarray of claim 48, wherein at least 95% of the probes on the microarray are present in any one of Tables 1-6.

54. The microarray of claim 48, wherein at least 98% of the probes on the microarray are present in any one of Tables 1-6.

55. A microarray for distinguishing ER(+) and ER(-) cell samples comprising a positionally-addressable array of polynucleotide probes bound to a support, said polynucleotide probes comprising a plurality of polynucleotide probes of different nucleotide sequences, each of said different nucleotide sequences comprising a sequence complementary and hybridizable to a different gene, said plurality consisting of at least 20 of the genes corresponding to the markers listed in Table 1 or Table 2, wherein at least 50% of the probes on the microarray are present in Table 1 or Table 2.

56. A microarray for distinguishing *BRCA1*-related and sporadic cell samples comprising a positionally-addressable array of polynucleotide probes bound to a support, said polynucleotide probes comprising a plurality of polynucleotide probes of different nucleotide sequences, each of said different nucleotide sequences comprising a sequence complementary and hybridizable to a different gene, said plurality consisting of at least 20 of the genes corresponding to the markers listed in Table 3 or Table 4, wherein at least 50% of the probes on the microarray are present in Table 3 or Table 4.

57. A microarray for distinguishing cell samples from individuals having a good prognosis and cell samples from individuals having a poor prognosis, comprising a positionally-addressable array of polynucleotide probes bound to a support, said polynucleotide probes comprising a plurality of polynucleotide probes of different nucleotide sequences, each of said different nucleotide sequences comprising a sequence complementary and hybridizable to a different, said plurality consisting of at least 20 of the genes corresponding to the markers listed in Table 5 or Table 6, wherein at least 50% of the probes on the microarray are present in Table 5 or Table 6.

58. A kit for determining whether a sample contains a *BRCA1* or sporadic mutation, comprising at least one microarray comprising probes to at least 20 of the genes corresponding to the markers listed in Table 3, and a computer readable medium having recorded thereon one or more programs for determining the similarity of the level of nucleic acid derived from the markers listed in Table 3 in a sample to that in a *BRCA1* pool and a sporadic tumor pool, wherein the one or more programs cause a computer to perform a method comprising computing the aggregate differences in expression of each marker between the sample and *BRCA1* and the aggregate differences in expression of each marker between the sample and sporadic pool, or a method comprising determining the correlation of expression of the markers in the sample to the expression in the *BRCA1* and sporadic pools, said correlation calculated according to Equation (3).

59. A kit for determining the ER-status of a sample, comprising at least one microarray comprising probes to at least 20 of the genes corresponding to the markers listed in Table 1, and a computer readable medium having recorded thereon one or more programs for determining the similarity of the level of nucleic acid derived from the markers listed in Table 1 in a sample to that in an ER(-) pool and an ER(+) pool, wherein the one or more programs cause a computer to perform a method comprising computing the aggregate differences in expression of each marker between the sample and ER(-) pool and the aggregate differences in expression of each marker between the sample and ER(+) pool, or a method comprising determining the correlation of expression of the markers in the sample to the expression in the ER(-) and ER(+) pools, said correlation calculated according to Equation (3).

60. A kit for determining whether a sample is derived from a patient having a good prognosis or a poor prognosis, comprising at least one microarray comprising probes to at least 20 of the genes corresponding to the markers listed in Table 5, and a computer readable medium having recorded thereon one or more programs for determining the similarity of the level of nucleic acid derived from the markers listed in Table 5 in a sample to that in a pool of samples derived from individuals having a good prognosis and a pool of samples derived from individuals having a poor prognosis, wherein the one or more programs cause a computer to perform a method comprising computing the aggregate differences in expression of each marker between the sample and the good prognosis pool and the aggregate differences in expression of each marker between the sample and the poor prognosis pool, or a method comprising determining the correlation of expression of the

markers in the sample to the expression in the good prognosis and poor prognosis pools, said correlation calculated according to Equation (3).

5

10

15

20

25

30

35

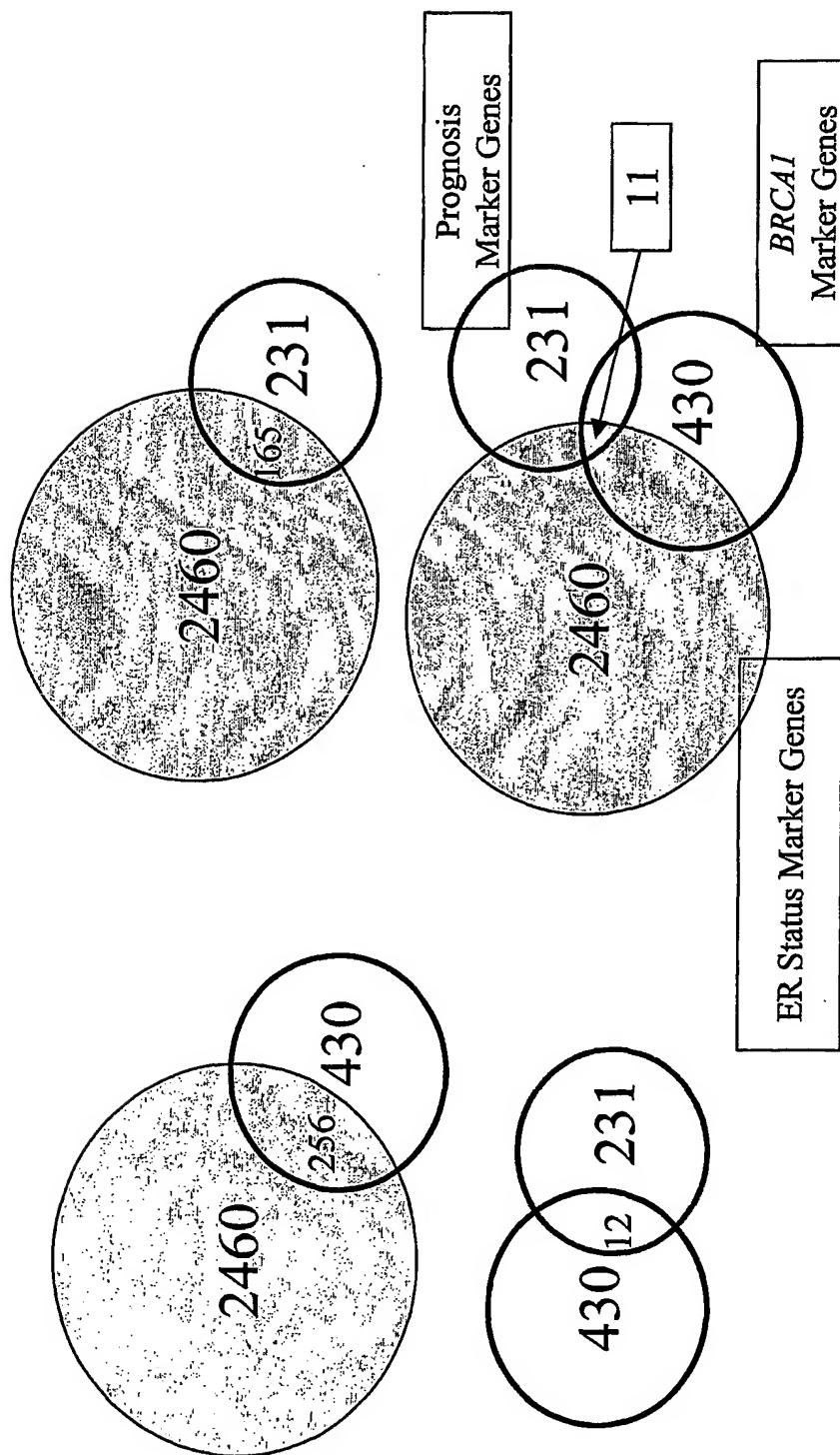


FIG. 1

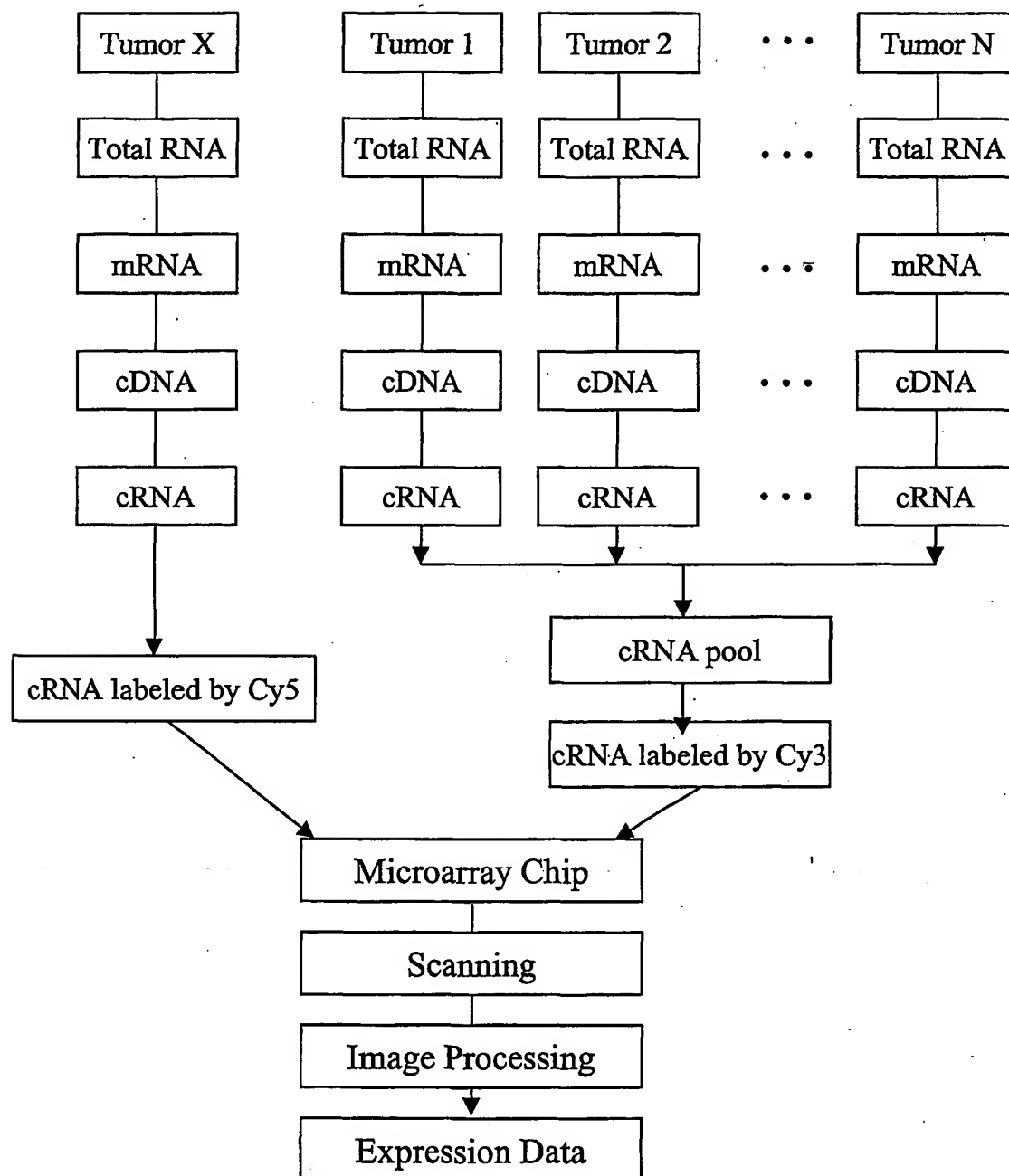


FIG. 2

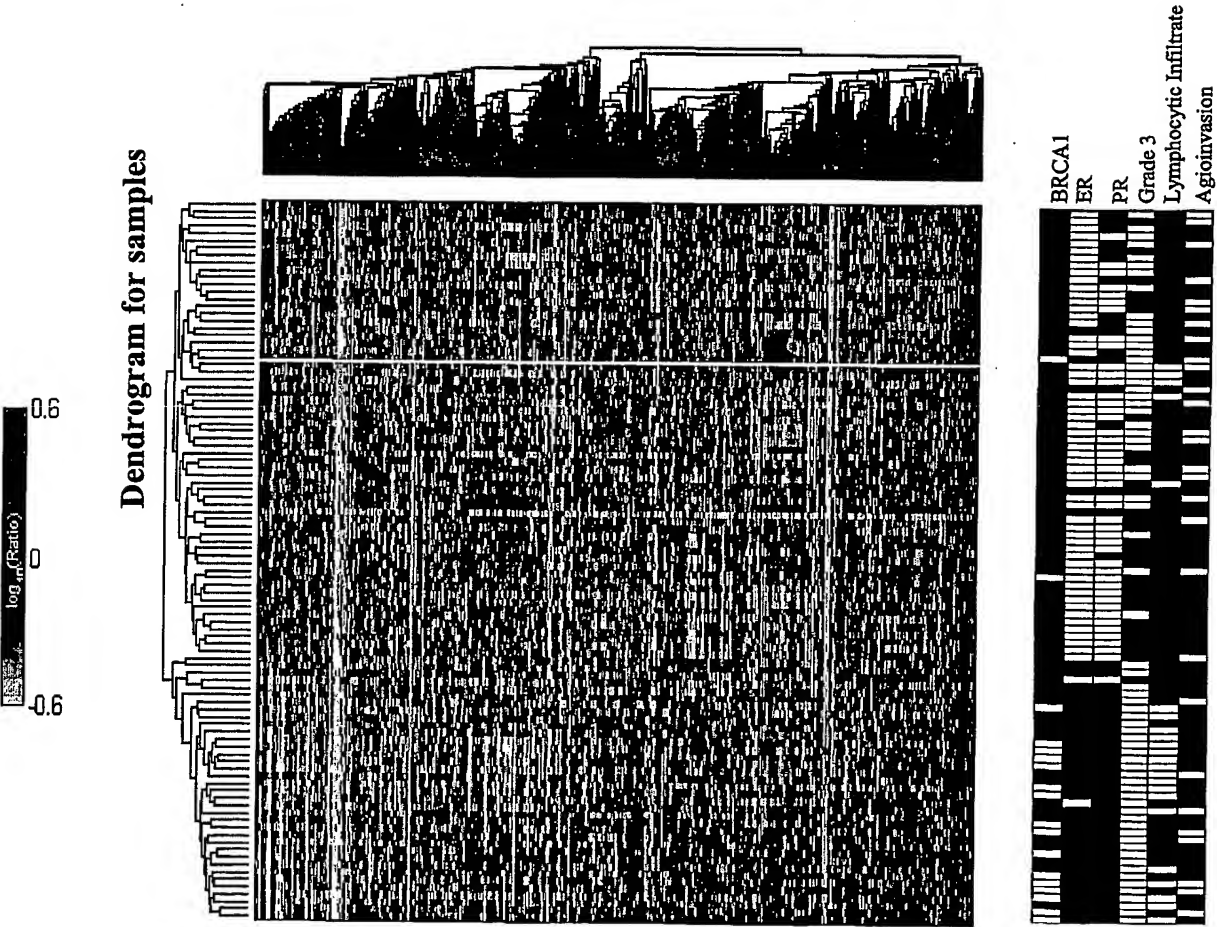


FIG. 3

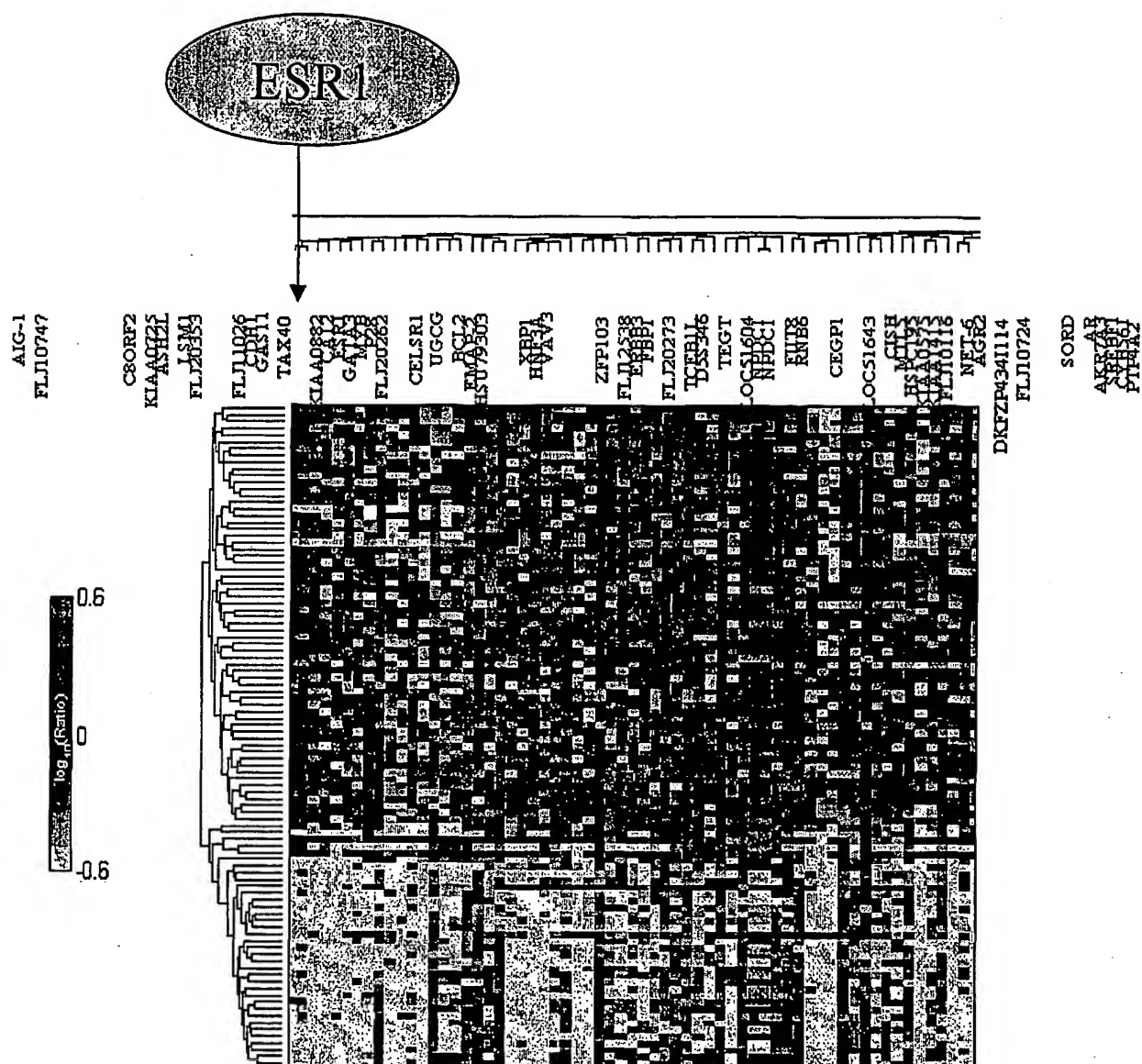


FIG. 4

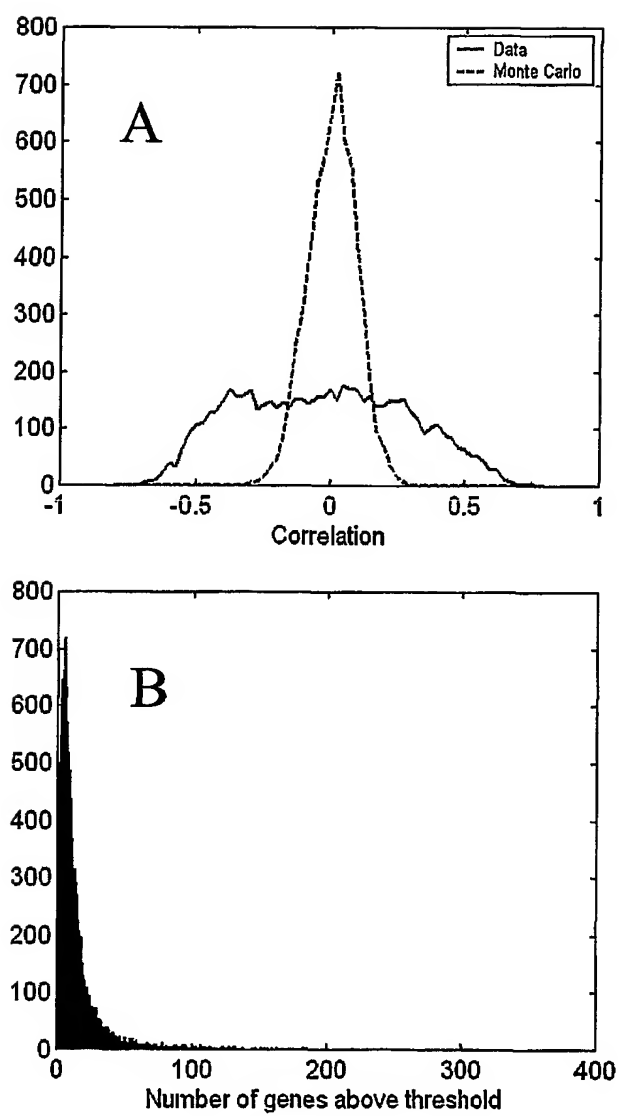


FIG. 5

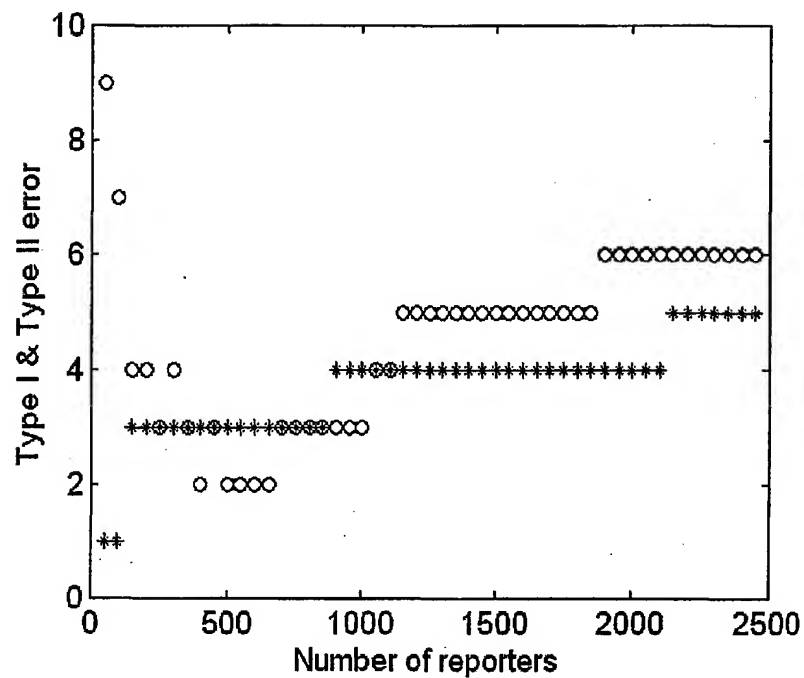


FIG. 6

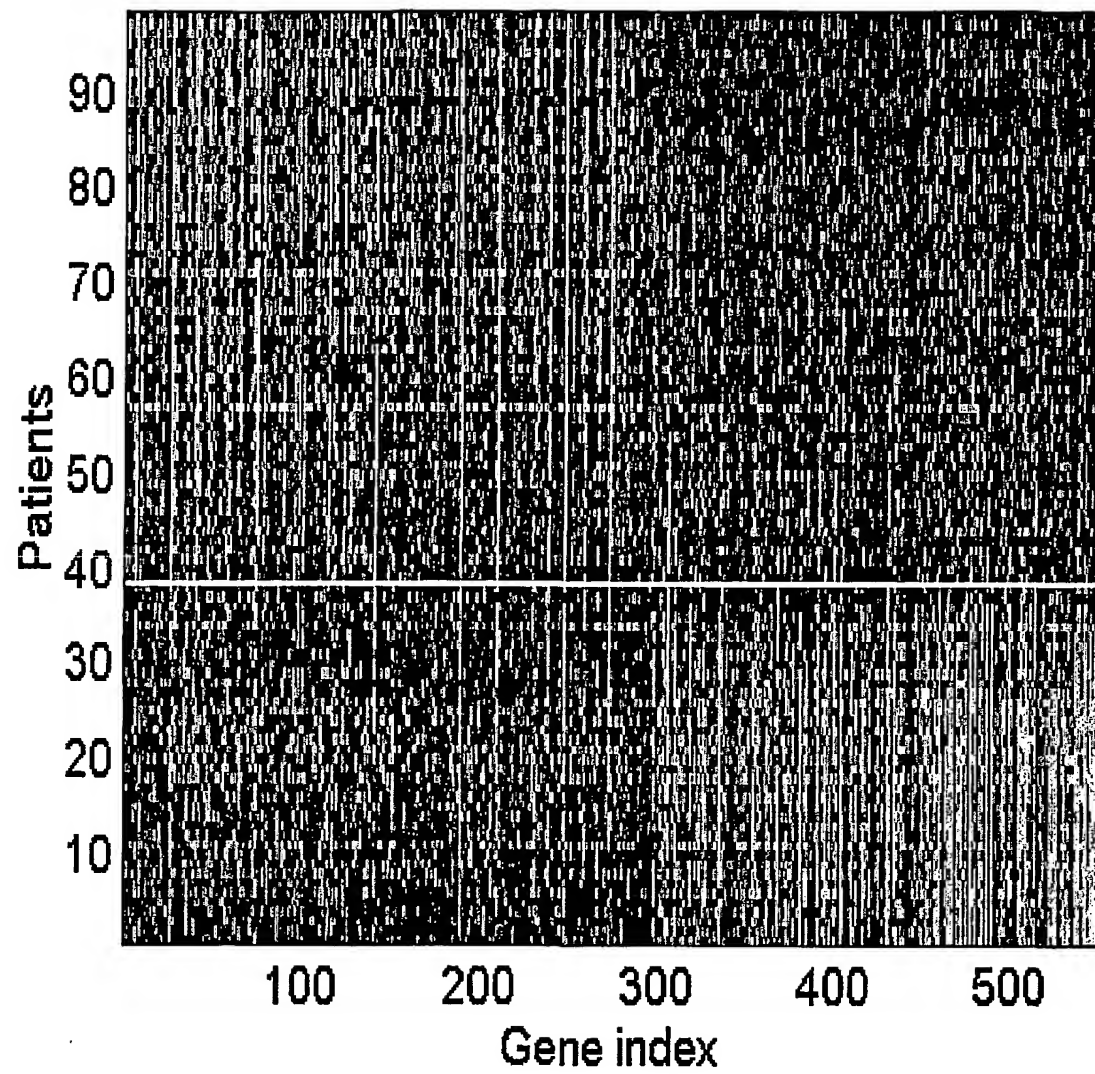


FIG. 7

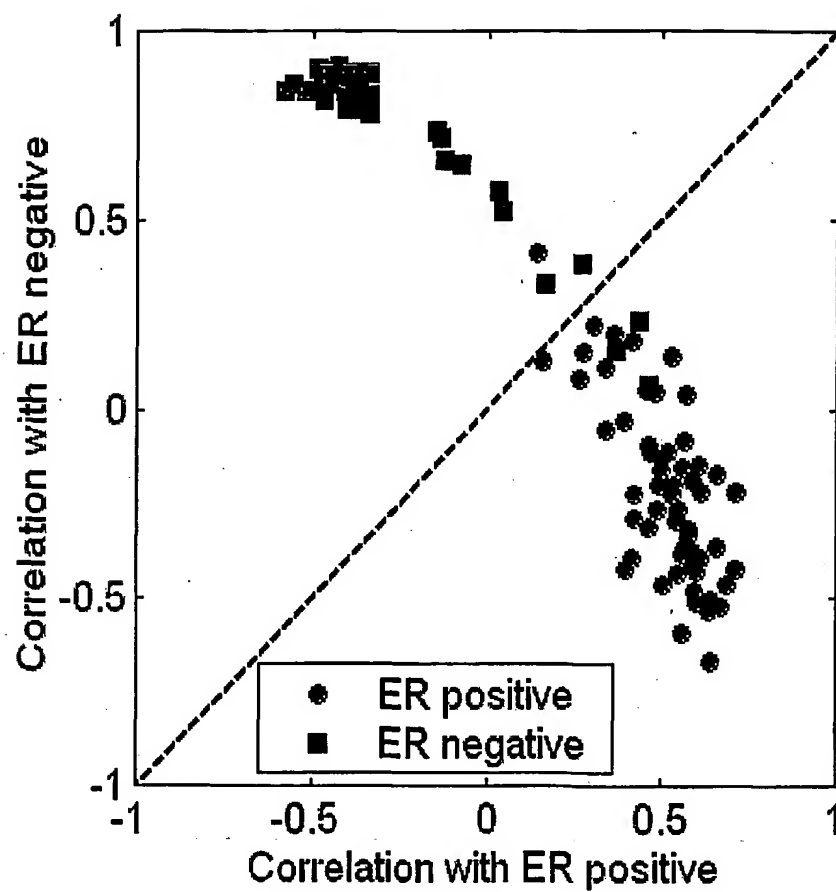


FIG. 8

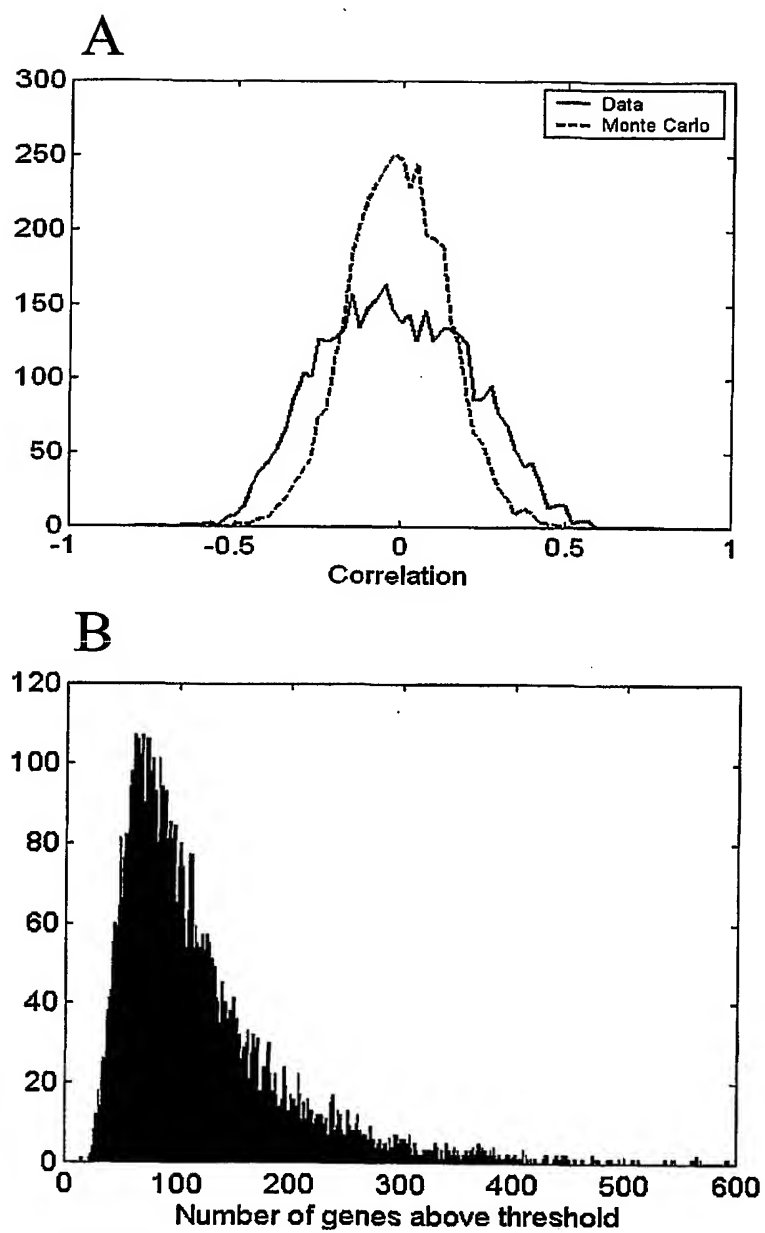


FIG. 9

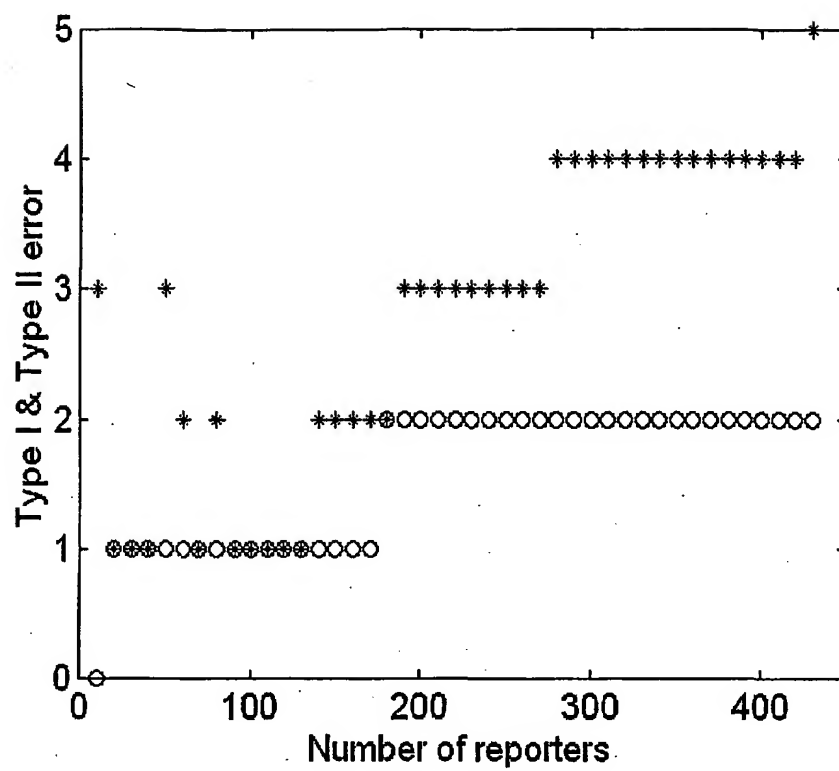


FIG. 10

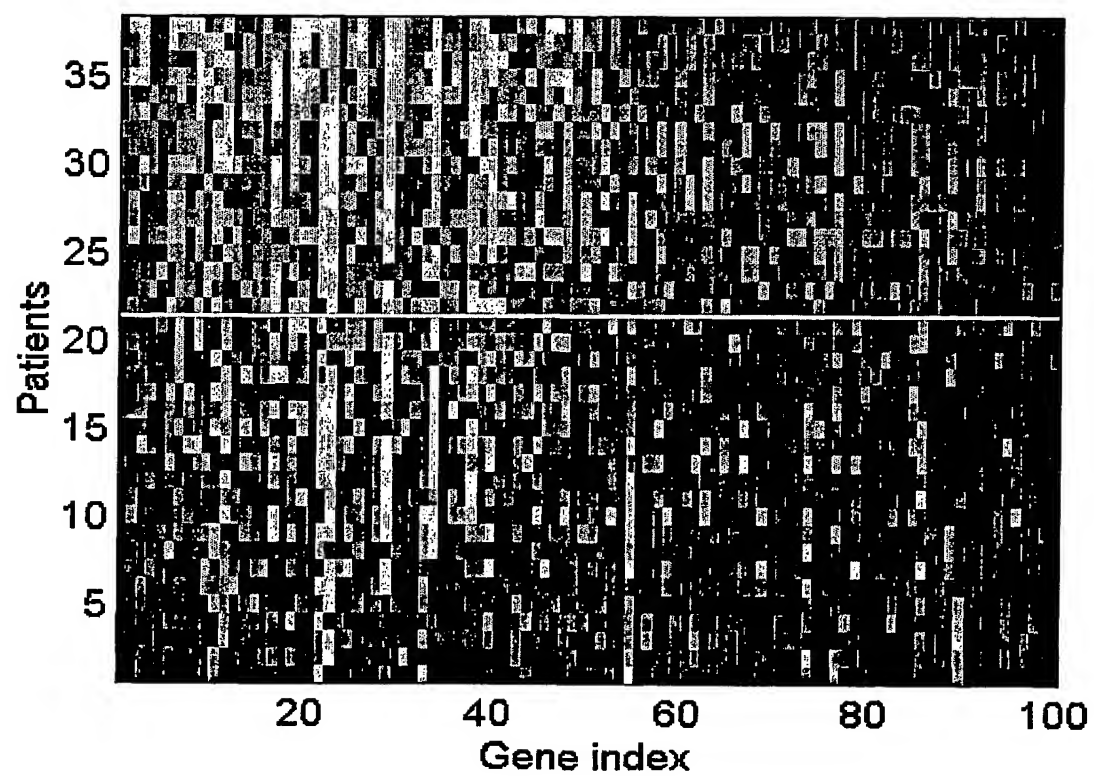


FIG. 11A

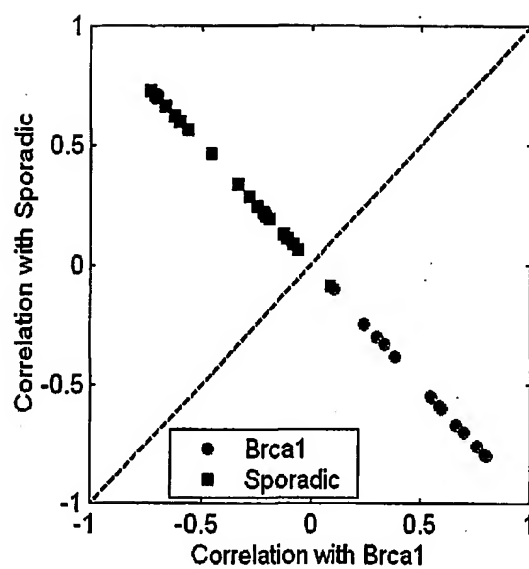


FIG. 11B

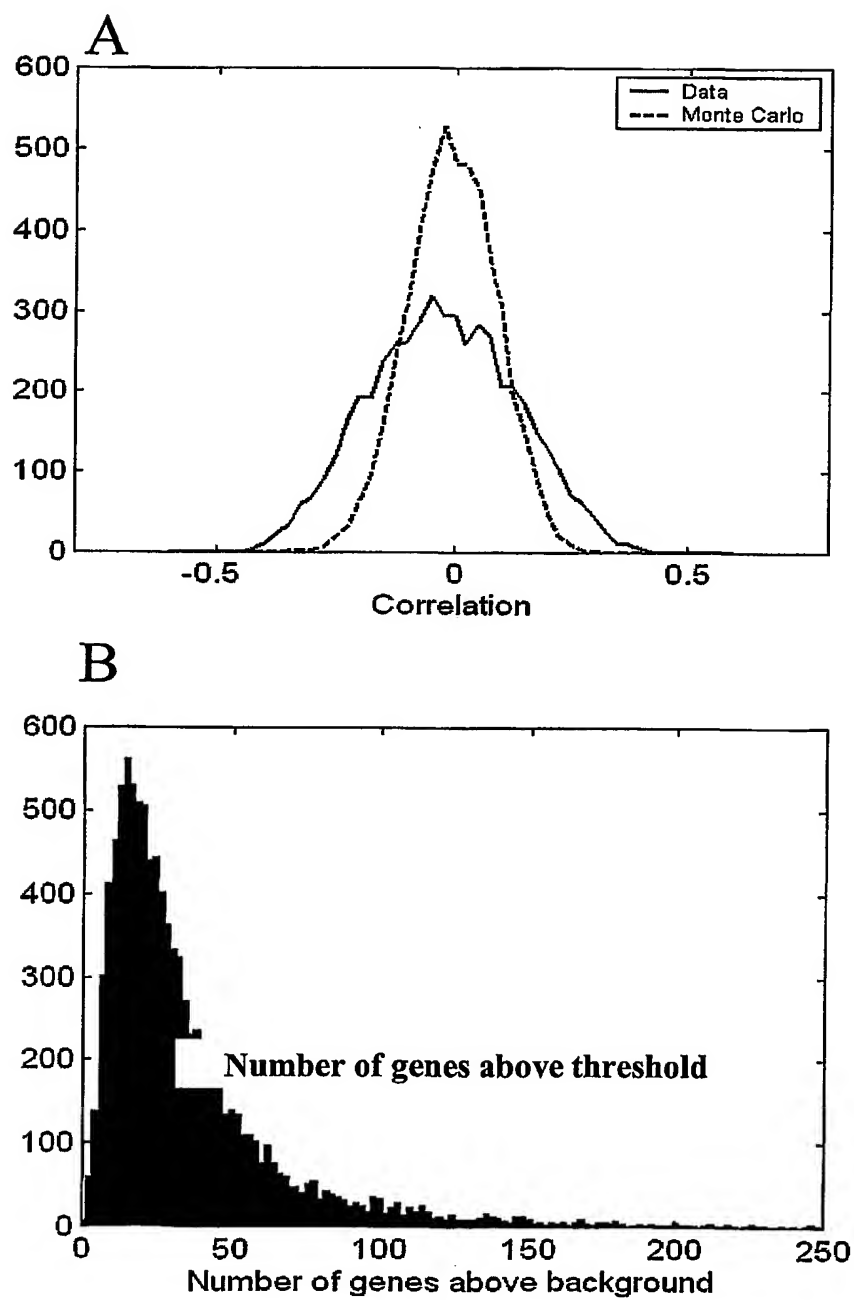


FIG. 12

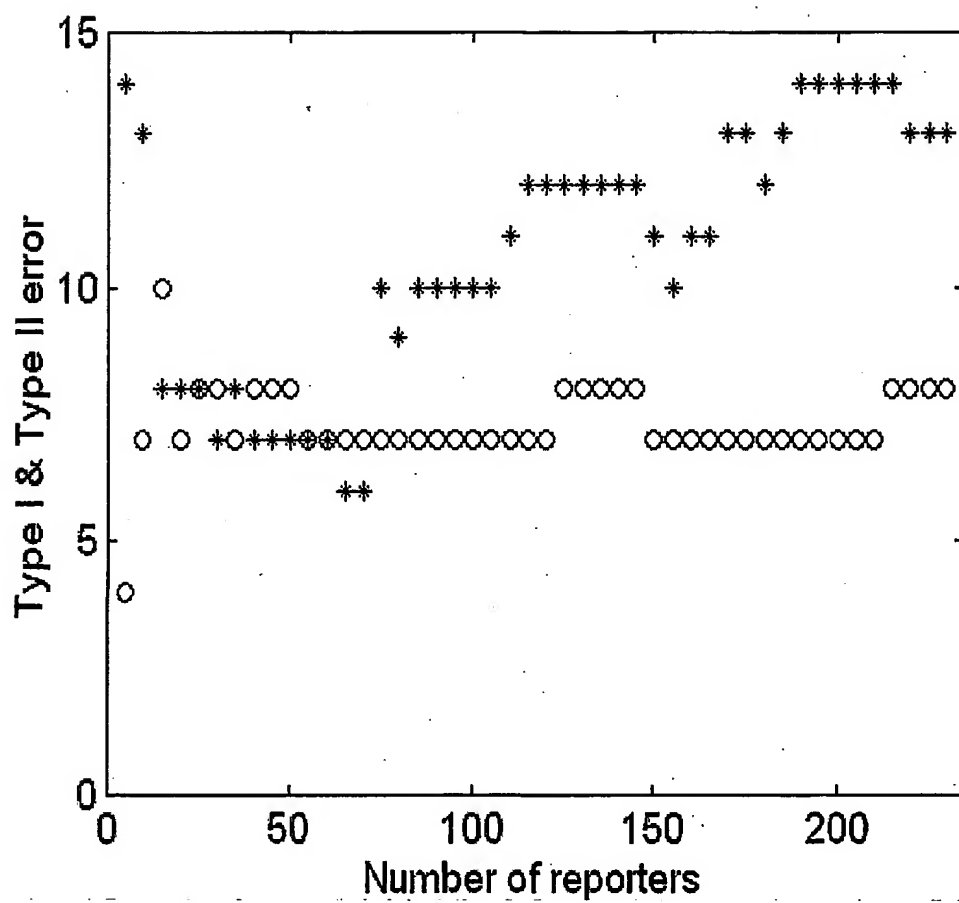


FIG. 13

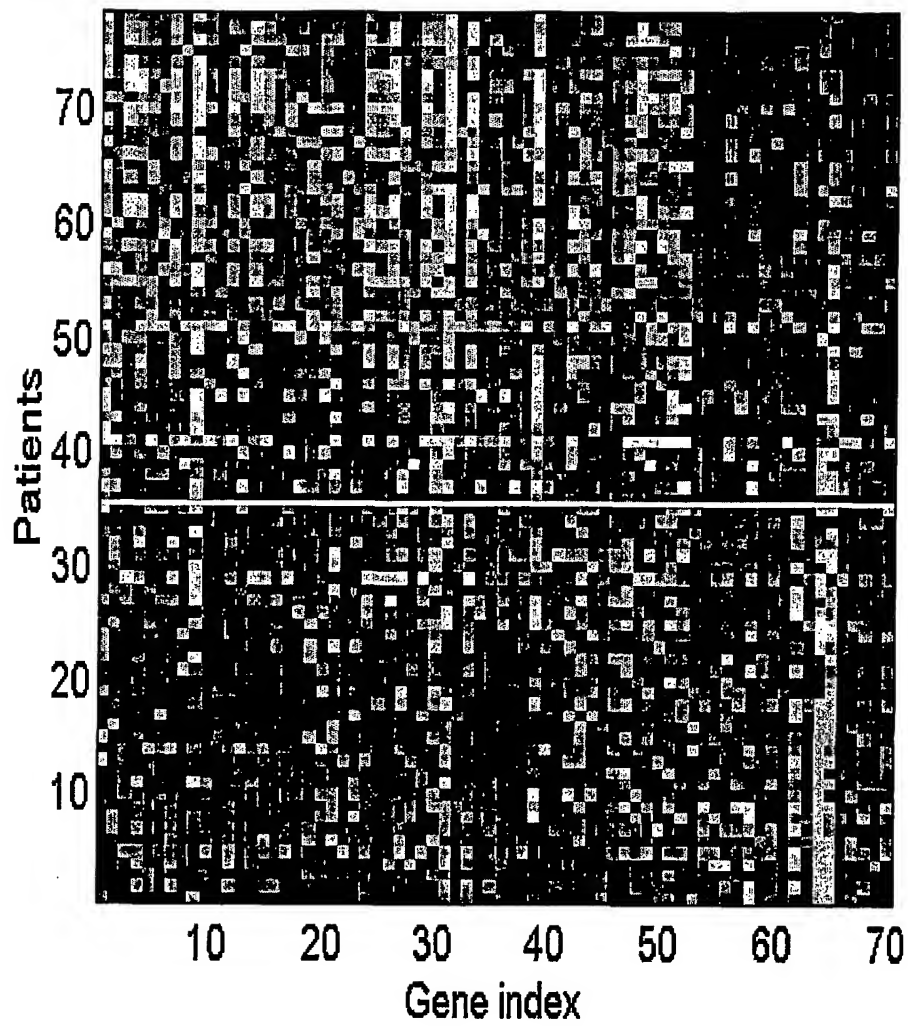


FIG. 14

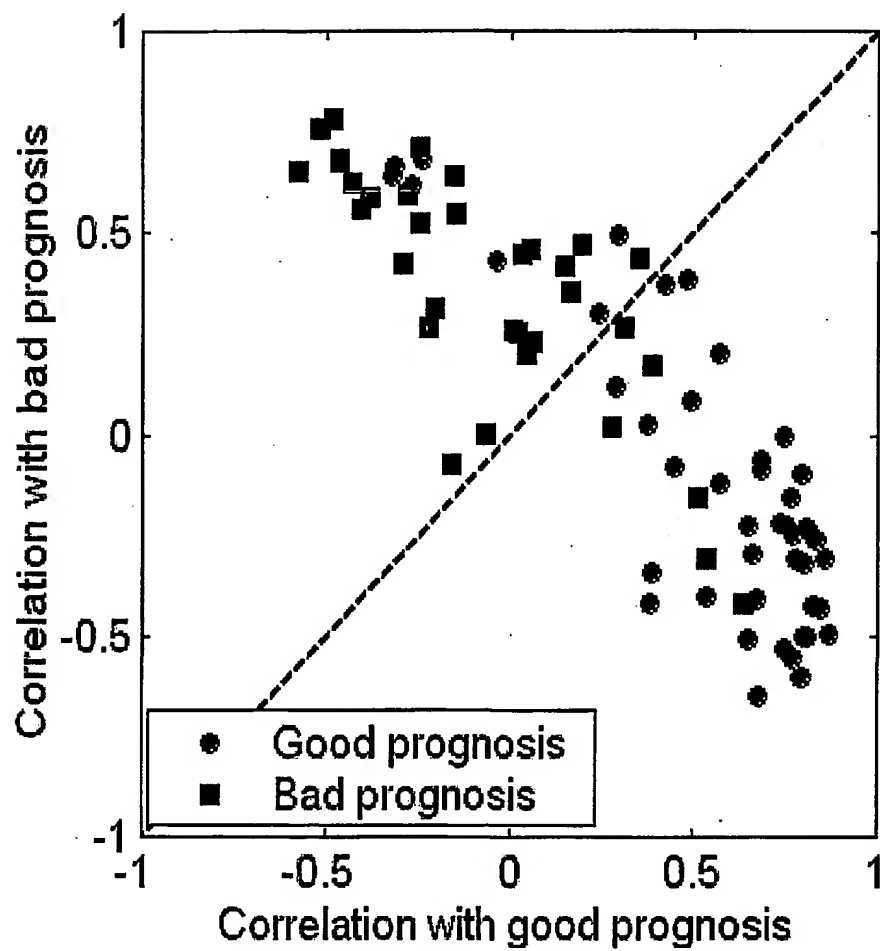


FIG. 15

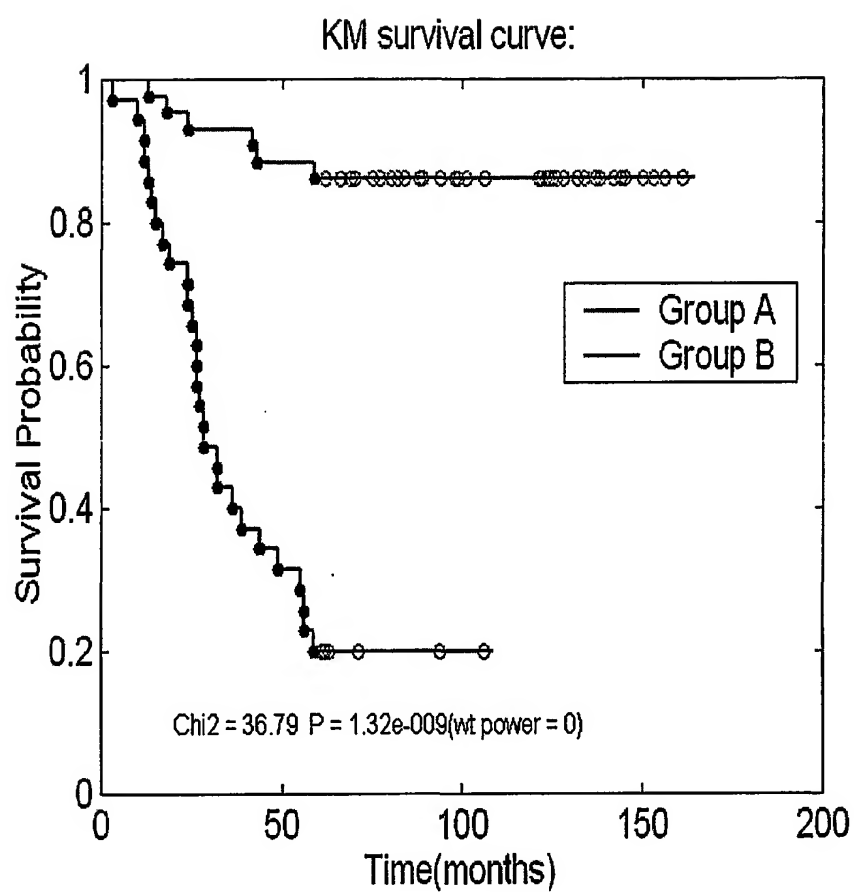


FIG. 16

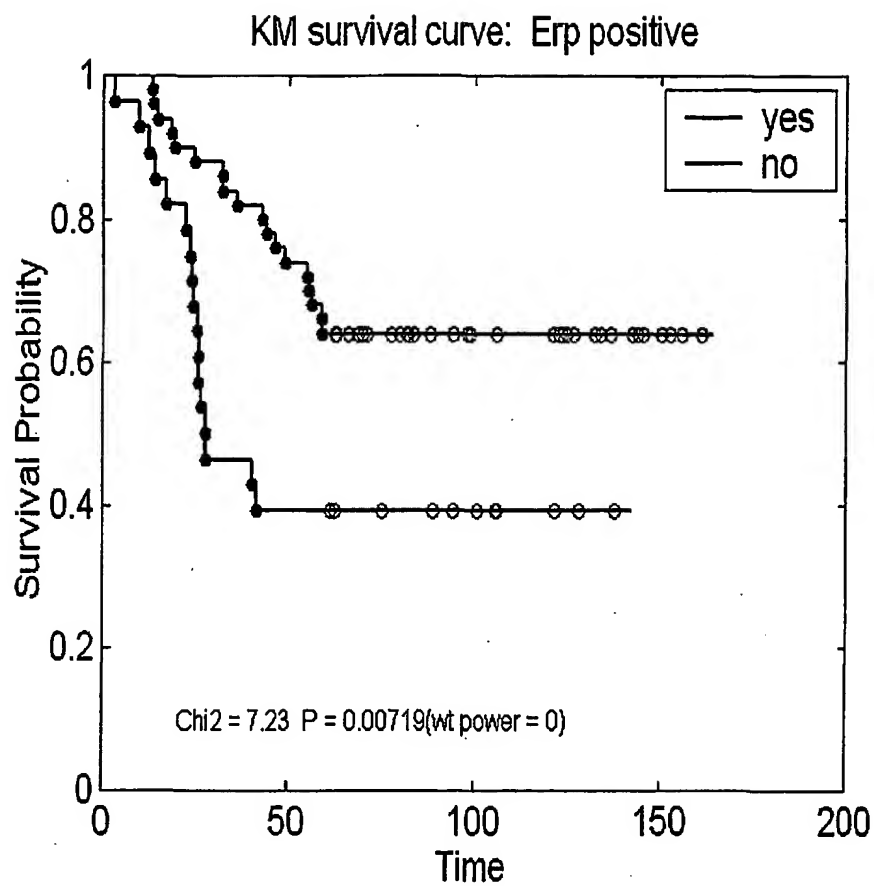


FIG. 17

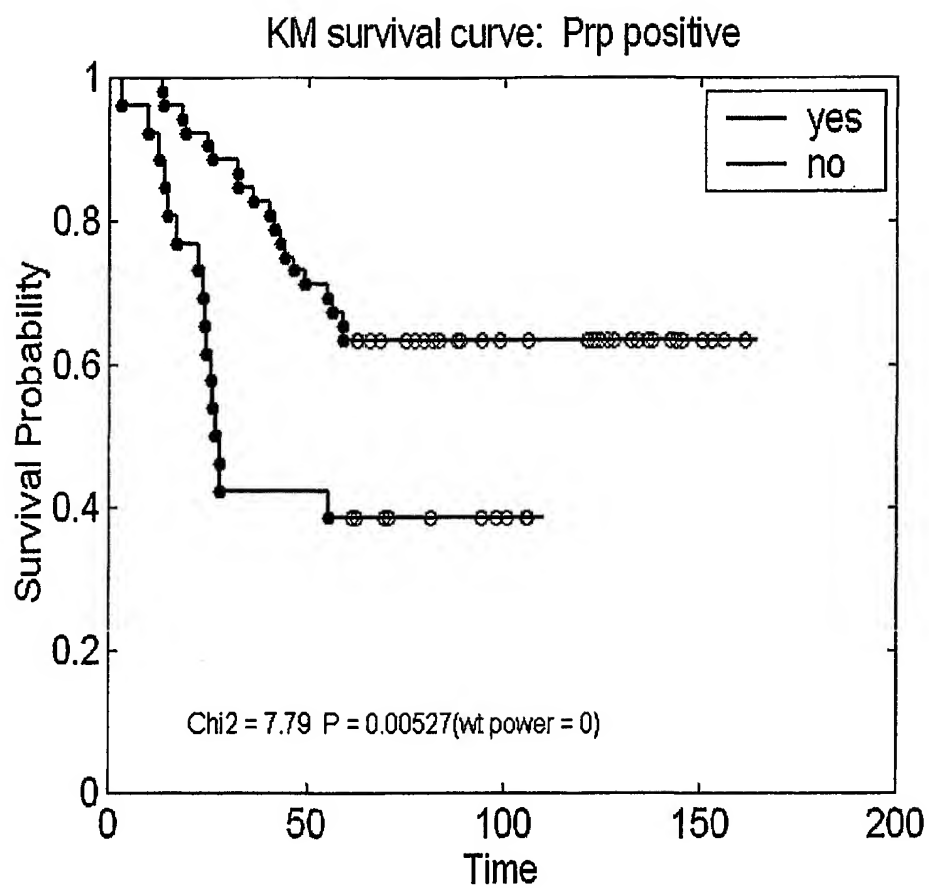


FIG. 18

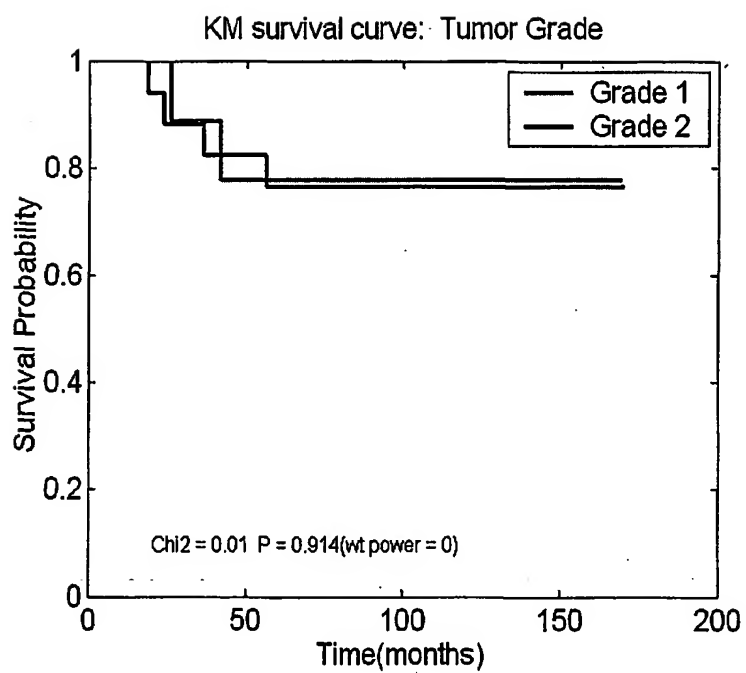


FIG. 19A

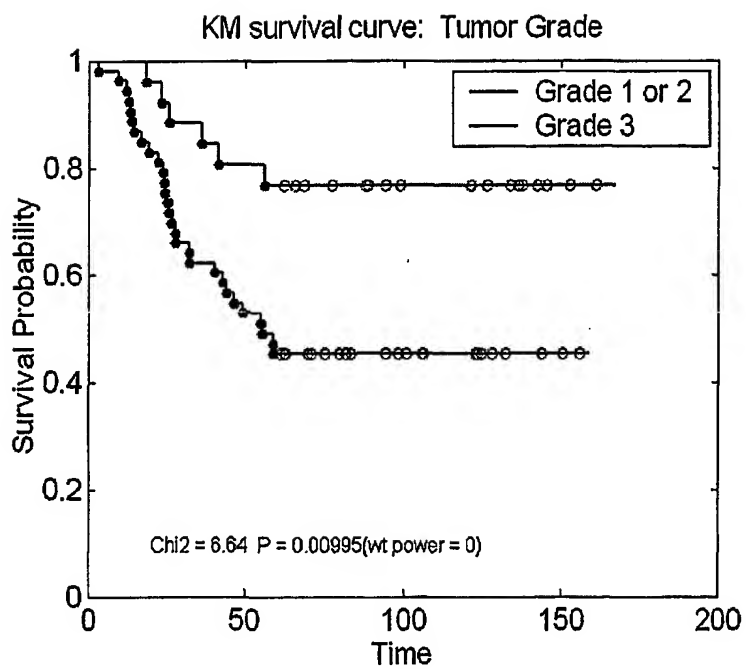


FIG. 19B

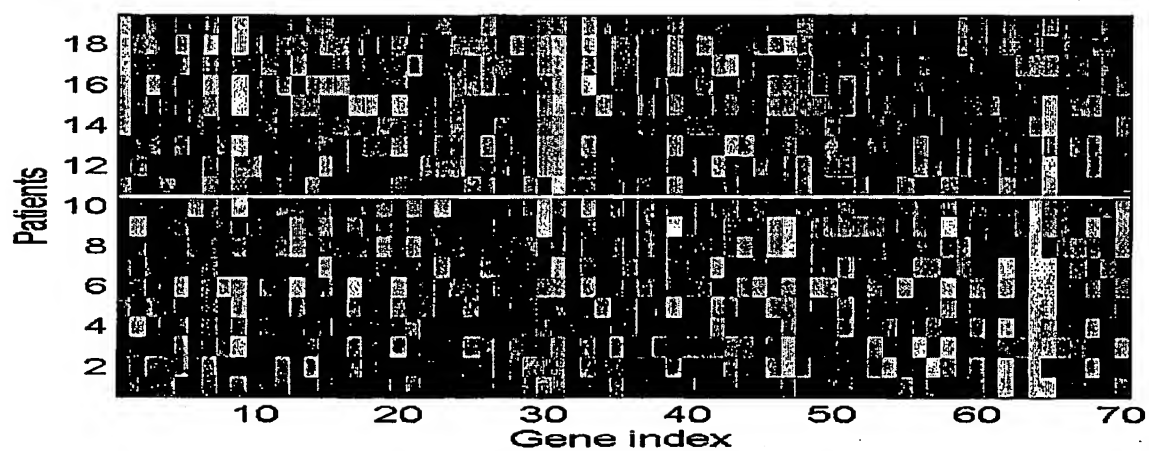


FIG. 20A

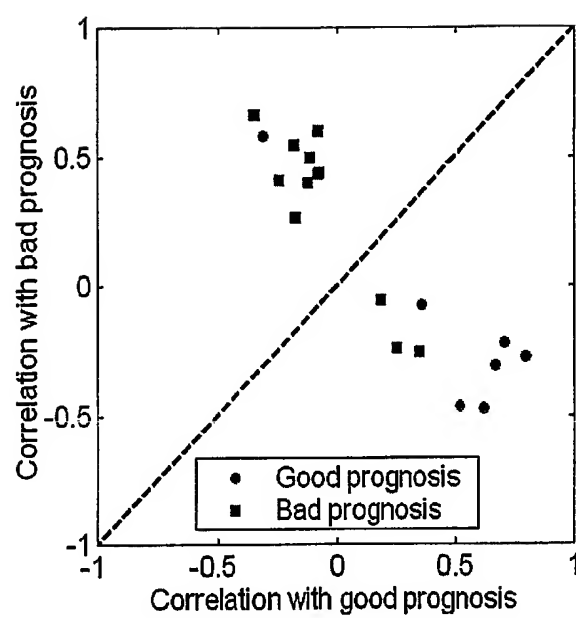


FIG. 20B

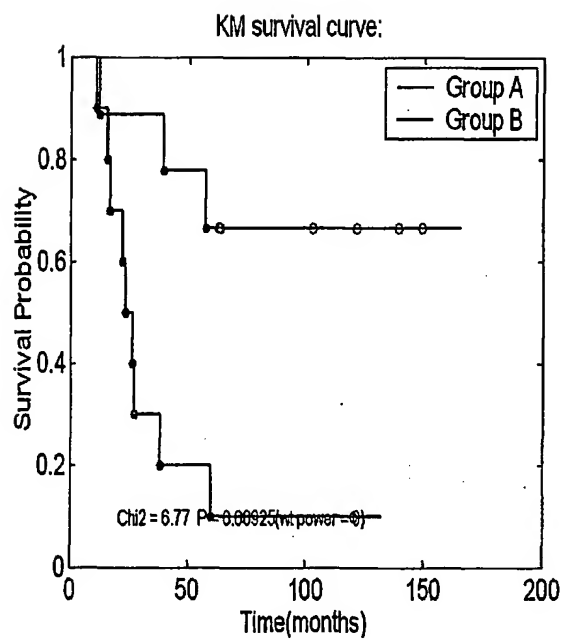


FIG. 20C

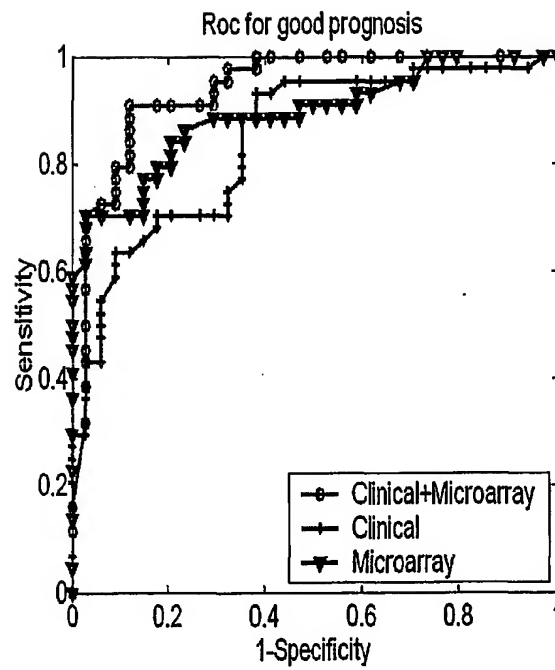


FIG. 21A

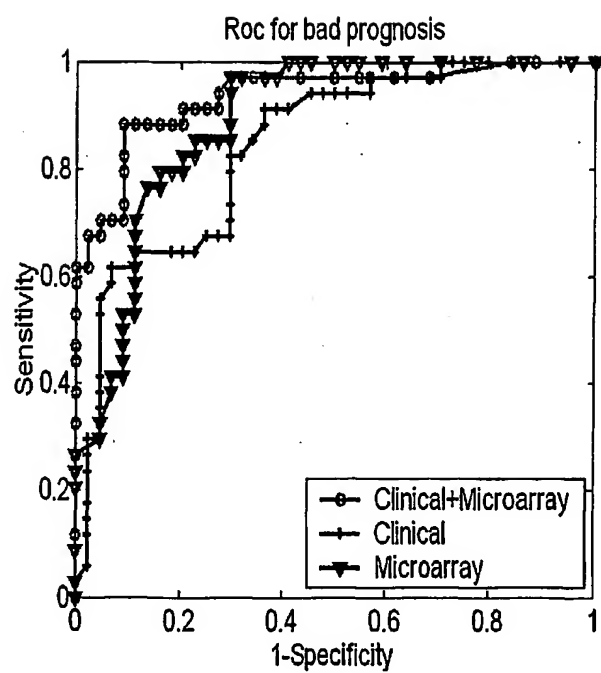


FIG. 21B

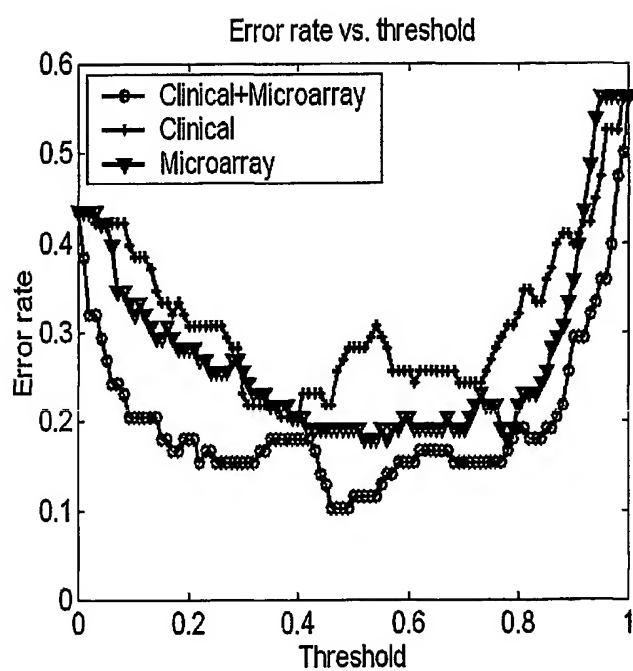


FIG. 21C

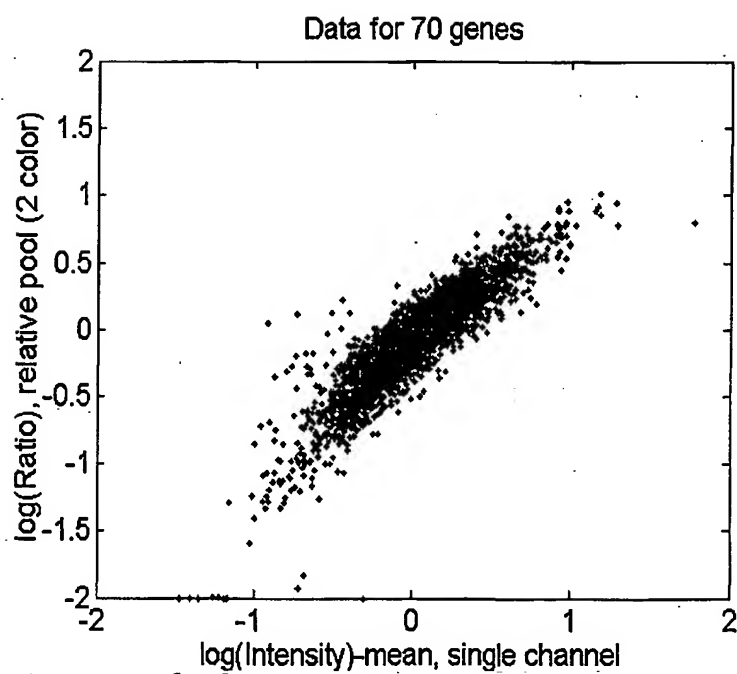


FIG. 22

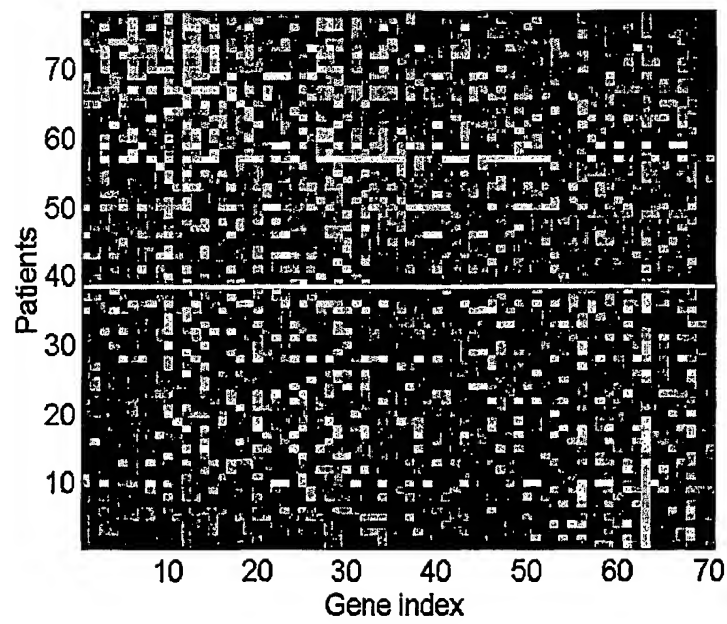


FIG. 23A

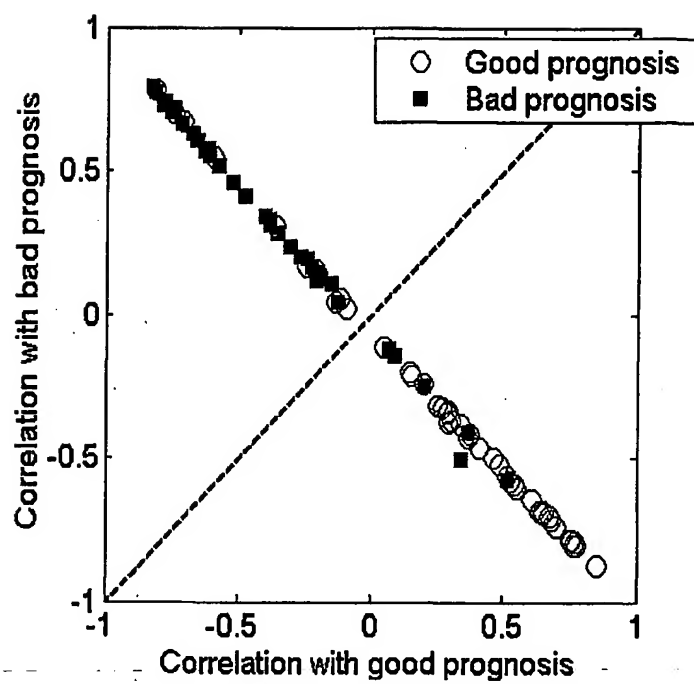


FIG. 23B

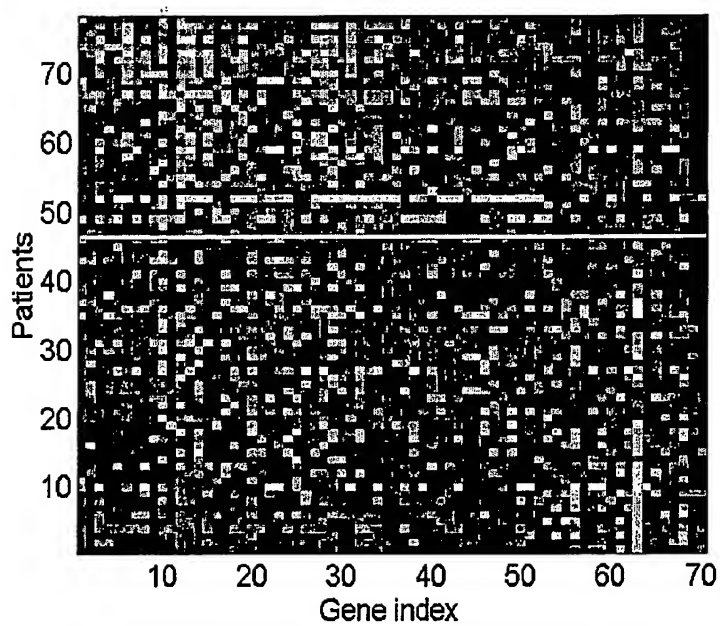


FIG. 24A

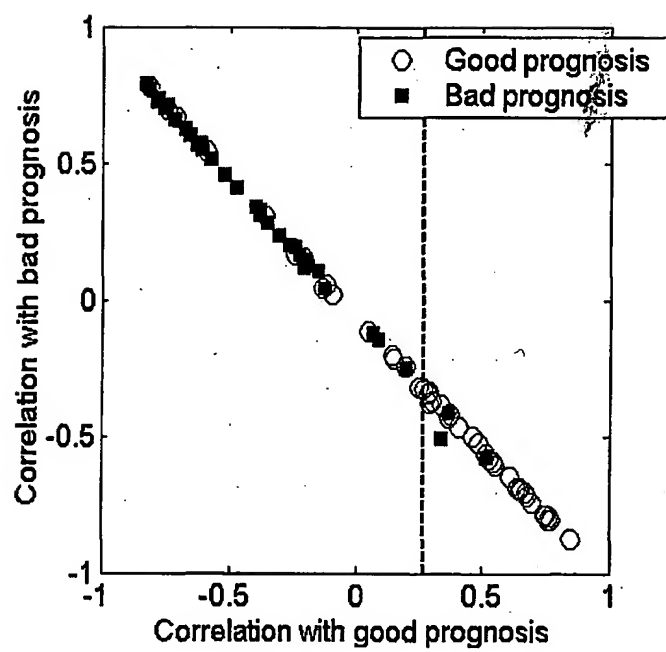


FIG. 24B

THIS PAGE BLANK (USPTO)

**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ **BLACK BORDERS**
- ☐ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- ☐ **FADED TEXT OR DRAWING**
- ☐ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- ☐ **SKEWED/SLANTED IMAGES**
- ☐ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- ☐ **GRAY SCALE DOCUMENTS**
- ☐ **LINES OR MARKS ON ORIGINAL DOCUMENT**
- ☐ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- ☐ **OTHER:** _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.

THIS PAGE BLANK (USP1U)